

Shooting the messenger

The Intergovernmental Panel on Climate Change has a creditable record of developing a scientific consensus and delivering it to policy-makers. What its critics really object to are the facts.

The drumbeat of diversionary noise that US coal and oil producers have created around climate-change science has a certain inevitable rhythm, all of its own. Take this year's events, for example (although those of any one of the past 15 years would do). The Bush administration, short of scientific advisers of its own but keen to justify its premature dismissal of the Kyoto Protocol on climate change, asked the National Academy of Sciences to conduct a rapid-fire study into climate-change questions.

The study's main charge was to assess the scientific evidence that greenhouse-gas emissions are contributing to global warming. To nobody's surprise, it answered in the affirmative. However, the final part of the study's charge asked: do the summary documents prepared by the Intergovernmental Panel on Climate Change (IPCC) reflect its underlying study of climate-change issues? The academy found that they did. Indeed, during the study, the academy asked US participants in the IPCC process if they felt that their views had been reflected fairly in the international organization's summary documents. Every participant who responded said that the summaries had been fair.

What a surprise, then, to find that the same charges of distortion-by-summary that climate-change sceptics made against the IPCC are now being repeated against the all-American National Academy. Within days of the study's release, critics were busy fulminating against the academy's process and the way in which the report had been summarized and represented in the media.

Inclusion tactics

The perennial charge against the IPCC (see News Feature, page 112) is that its summary-making process involves politicians and non-governmental organizations, as well as scientists. But in this regard, the IPCC process is actually ahead of that commonly used by the US National Academies and other groups engaged in advising governments around the world on assessing and managing risk.

A key element to producing a convincing scientific-risk assessment, according to the National Academy's seminal 1996 report, *Understanding Risk: Informing Decisions in a Democratic Society*, is that interested parties should be involved in the assessment while it is being made. The report argued that this is preferable to the traditional, linear model of risk management, whereby scientists assess the risk and then dump their findings on the politicians who are expected to manage it.

So the IPCC process is a markedly modern one whose overall effect has been, on balance, to moderate the findings of its various working groups. Detailed examination of the process shows that it has a reasonable track record of producing summary documents that, while lacking all of the detailed qualification contained in the full working-group reports, honestly reflect their findings.

Why, then, is it embroiled in so much criticism? If the IPCC's record is set against its critics' charges, it soon becomes apparent that the answer to this question lies in the latter group's unscrupulous determination to defy the facts on climate change to the bitter end.

Right from the outset, the approach of certain industrial lobby groups in the United States has been to resist, resist and resist

again the mounting evidence that the consumption of fossil fuels is producing emissions that change the make-up of the atmosphere and may endanger the future of the planet. The industry groups in question are accustomed to the untrammelled purchase of political power in the United States and have consistently sought to distort the climate-change debate for their own purposes.

Creating dissent

To this end, they have championed specious scientific findings and worked to establish a bogus scientific debate between their own 'experts' — many of whom are not even atmospheric scientists — and the consensus view of climate researchers. In doing this, they have deliberately set out to take maximum advantage of media gullibility, ensuring that stories on the problem include both 'sides' of the debate.

Science can only progress through its strong tradition of debate and dissent. But this particular debate discredits the notion of scientific dissent. Some of the climate-change dissidents bring to mind the AIDS dissidents who spent the 1990s putting about notions that HIV didn't cause AIDS and that there was no AIDS pandemic in Africa. Others resemble the tobacco-industry dissidents who resisted that industry's regulation to the last, on the basis of what later emerged as the misrepresentation of science funded by cigarette manufacturers.

The campaign to confuse and delude the US public on global warming has had its successes, but it has been less effective than its originators like to pretend. The public is not amused by President Bush's dismissal of the Kyoto Protocol and it is notable that Senator John McCain (Republican, Arizona) returned from his unsuccessful fight with Bush for the Republican nomination convinced, by his encounters with the public, that action on carbon emissions was needed.

The Bush administration persists in repeating the fiction, both at home and abroad, that the US Senate has already rejected the Kyoto Protocol by 95 votes to 0. In fact, on 25 July 1997, the Senate passed the non-binding Byrd–Hagel Resolution by that margin. This called for developing countries to be more involved in the protocol, which was negotiated a few months later.

Senator Robert Byrd (Democrat, West Virginia), the co-sponsor of that resolution, lambasted the Bush administration on 4 May of this year for its summary rejection of the Kyoto process, and many US senators do, in fact, support action — even mandatory action — on carbon emissions. The political consensus in the United States is that Bush's ratings with the public have largely fallen on account of his environmental missteps, of which the abandonment of Kyoto was the most celebrated.

The Senate is by no means ready to ratify the Kyoto Protocol in its existing form, but neither is US public opinion where Bush pretends it to be, in outright opposition. No amount of sniping at the IPCC by the Global Climate Coalition — itself a depleted operation following the defection of the car manufacturers and some oil companies — will convince the US public that global warming is an imaginary problem. The IPCC should continue its valuable work in the knowledge that its integrity and adherence to the facts will ultimately prevail, bringing credit to both its architects and its participants. ■





No frontiers
Europe and China
embark on cosmic
collaboration
p106



End of the line
Compromise over
stem cells gets
thumbs down
p107



Sinking fast
Doubts cast over
trees' ability to
sequester CO₂
p108



Crop crusade
UN flags benefits of
transgenic crops for
developing world
p109

Genetics group targets disease markers in the human sequence

David Adam, London

Researchers are planning a new public-private consortium to map genetic variation across the entire human genome.

The proposed collaboration between several biotechnology companies and public laboratories, including the Sanger Centre in Britain and the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, aims to produce a publicly available genetic map of some 300,000 'haplotypes' within two years.

Haplotypes are sets of genetic markers that are close enough on a chromosome to be inherited together. They can be used to unravel the genetic differences that make some people more susceptible than others to conditions such as diabetes and heart disease.

Plans for the \$60-million project will be discussed at a meeting in Washington on 18–19 July. The US National Institutes of Health and the Wellcome Trust, the UK-based charity that funds the Sanger Centre, are keen to support the consortium, and several pharmaceutical companies have also been approached for possible funds.

"The objective is to create a framework map across the entire genome that will be of great use in disease-gene finding, in pharmacogenetics and in evolutionary studies," says Lon Cardon, head of bioinformatics at the Wellcome Trust Centre for Human Genetics in Oxford.

Cardon says the proposed consortium would build on both the human genome sequencing project and the recently completed effort to identify single-nucleotide polymorphisms (SNPs). The latter collaboration identified some two million SNPs — sites where the genetic code between individuals differs by just one base.

"Now we have the potential to actually do something with those variations by seeing how they are associated across the genome," Cardon says.

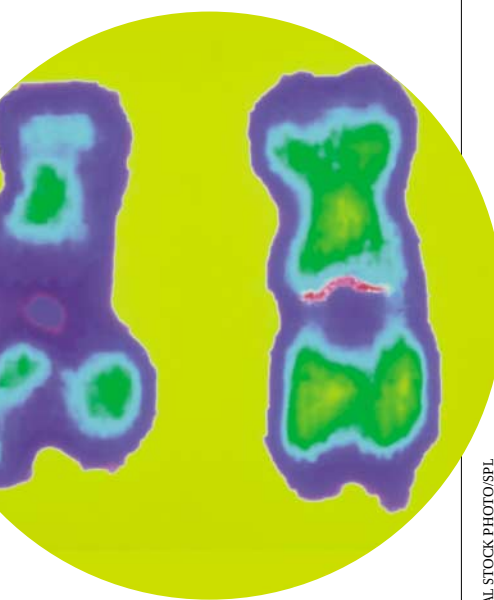
SNPs act as markers for genetics researchers sifting the genome for disease-causing genes. They allow genes to be traced through generations, for example, and can flag up genetic differences between people affected and unaffected by a disease.

But using SNPs alone can be difficult and expensive, partly because it is hard to trace individual SNPs in a genome containing three billion base pairs. Using haplotypes eases some of these difficulties, as each contains a group of SNPs that tend to be inherited together. So the two million or so separated markers are consolidated into more manageable clusters, making it easier to identify variation in the genome.

Haplotypes are found by analysing genotype data. The new consortium would essentially be a high-throughput genotyping effort, in much the same way that the public human genome project is a high-throughput sequencing effort.

The consortium would produce a reference map of haplotype regions across the genome. Researchers studying diabetes, for example, would then be able to see which haplotype regions differ in affected individuals, helping to narrow down the region of the genome associated with the condition.

"It's more informative if you can link two markers together in the inheritance pattern than just use one marker or the other," says Mike Boyce-Jacino of Orchid BioSciences, a biotechnology firm based in Princeton, New



Missing link: the chromosome on the left lacks a gene (red) that may protect against cancer.

Jersey, and one of the companies involved. "What you're really trying to do is to reduce the background noise when you go in to look for a genetic association with the inheritance of a disease," he says. ■

Physicists show what really matters

Colin Macilwain, Washington

Researchers at the Stanford Linear Accelerator Center (SLAC) in California have confirmed an important tenet of the standard model of high-energy physics. They have successfully measured a parameter that may help to explain the preponderance of matter over antimatter in the Universe.

An international team of 600 physicists announced on 6 July that they had measured the parameter, which expresses the degree of asymmetry between heavy subatomic particles called B mesons and their corresponding anti-B mesons.

The long-awaited finding, based on the observation of 32 million decay events by a

1,200-tonne detector called BaBar, confirms the existence of the overall asymmetry, known as charge-parity (CP) violation.

"We're absolutely delighted," says Stewart Smith, a physicist at Princeton University and spokesman for the experiment. "We expected it would take another year to get this far."

The 'B Factory' experiment, which generated the result, was devised and built by a team led by Jonathan Dorfan, who is now director of SLAC (see *Nature* 403, 586; 2000).

A rival international team at the KEK particle physics laboratory in Japan is set to announce its own measurement of CP violation at a news conference next week. ■

Europe hooks up with China for space first

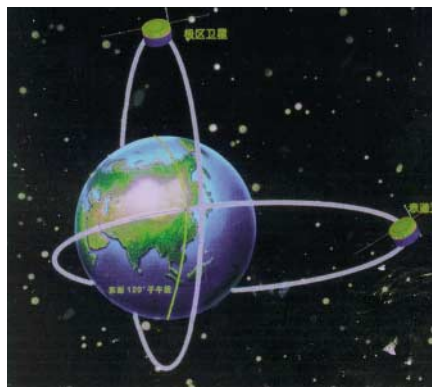
Sally Goodman, Paris

European space experiments will fly on board Chinese satellites for the first time in a joint project to study the magnetosphere — the magnetic shield that surrounds the Earth.

The mission, known as Double Star, was announced on 9 July by the European Space Agency (ESA) and the Chinese National Space Administration. It will involve two small Chinese satellites flying in orthogonal orbits — around the poles and the equator, respectively (see right) — to gather data on how the Sun affects the magnetosphere's behaviour.

In its first substantial collaboration in space science with China, the ESA will contribute 10 of the project's 18 experiments, costing about 8 million euros (US\$6.8 million). ESA officials are confident that the project will lead to more extensive collaborations with China in the future — despite American misgivings about sharing space technology with China.

Double Star is the latest of several projects designed to investigate the relationship between the Sun and the magnetosphere. Interaction between solar winds and the magnetosphere cause spectacular physical effects, including the aurora — or northern lights — seen over the polar skies, and massive magnetic storms that disrupt electricity supplies and radio communications.



Doubleplusgood: the orbits of the two satellites will allow investigation of the magnetosphere.

But many of these projects have suffered from technical glitches. Germany's Equator-S satellite, for example, was launched in 1997 but stopped sending back data after only five months when its batteries failed. And Cluster, a project involving the ESA, NASA and other partners, was lost in 1996 when its Ariane 5 rocket launcher failed.

But Cluster — a flotilla of four identical spacecraft — was relaunched last summer, and Michael Fehringer, an ESA scientist working on the project, says early data indicate that the mission "will revolutionize our understanding of the magnetosphere".

The ESA instruments that will fly on

Double Star are identical to those on Cluster, but they will fly in different orbits. To keep costs down, the equipment will mostly be put together using spare parts left over from Cluster. But because of US restrictions on exports to China, instruments of American origin will be rebuilt in Europe for the new mission. For example, André Balogh of Imperial College, London, says that his Double Star experiment will use a British-built magnetometer to replace the American one he used on Cluster.

The ESA hopes that the Double Star satellites, which are scheduled for launch on Chinese Long March 2C rockets in December 2002 and March 2003, will operate in parallel with the four existing Cluster spacecraft. Cluster is only funded until January 2003, but the Europeans hope to extend its life by at least two years.

Even if this does not happen, Double Star will allow researchers to gather new information, says Andrew Fazakerley of the Mullard Space Science Laboratory at University College London, who is leading one of the experiments. A process called magnetic reconnection — whereby particles are accelerated towards Earth's magnetic poles from its huge magnetic tail — will be observed in unprecedented detail from Double Star's equatorial orbit, he says.

♦ <http://www.esa.int>

Institutes prepare for pioneering bioinformatics work

Declan Butler, Paris

Two women scientists are to take control of two of Europe's leading bioinformatics initiatives, with a remit to expand intramural research activities.

Janet Thornton, a structural-biology professor at University College London, has been named as research director of the European Bioinformatics Institute (EBI) near Cambridge in Britain. She succeeds Michael Ashburner, who has co-directed the centre with Graham Cameron since it opened five years ago.

And Nadia Rosenthal, an associate molecular-biology professor at Harvard University, will succeed Klaus Rajewsky as head of the mouse biology programme at the European Mouse Mutant Archive near Rome.

Both facilities are affiliated to the European Laboratory of Molecular Biology (EMBL), and have until now acted chiefly as service providers to outside researchers. But Fotis Kafatos, director general of the EMBL, says that the laboratory is now shifting its emphasis to functional genomics, and he expects the facilities to play a more active role



Incoming: Janet Thornton (right) will head the EBI and Nadia Rosenthal will lead Europe's mouse mutant archive.

in pioneering original research of their own.

The EBI recently received a multimillion grant from the European Union (see *Nature* 411, 229; 2001), securing its main database programmes. Thornton says that the next step is to build a research environment around these projects. "To make progress in understanding whole systems we need to integrate these genomic, proteomic and other resources," she says.

Thornton's main research interests are in

the analysis and modelling of protein structures and in rational drug design. She says that she is keen to add 'chemoinformatics' to the EBI's traditional bioinformatics role and to strengthen the links between the EBI and the pharmaceutical industry.

Ashburner, who has split his time between the EBI and his faculty position at the University of Cambridge, will now return to full-time research at the university. But he says that he is keen to help Thornton settle in, and will remain involved with the EBI to oversee a large expansion of its gene ontology (GO) consortium, an effort to build controlled vocabularies that will assist in the cross-searching of biological databases (see *Nature* 411, 631–632; 2001).

Approaches such as GO are absolutely critical, says Thornton. "Without being able to describe function in a computer-readable form — without functional ontologies — we will not be able to describe whole systems."

♦ <http://www.ebi.ac.uk>

♦ <http://www.emma.rm.cnr.it>

♦ <http://www.geneontology.org>

Stem-cell fudge finds no favour with biologists

Meredith Wadman, Washington

A compromise answer to the hot political question of whether the US government should fund research on human embryonic stem cells has received a chilly response from cell biologists.

The suggested compromise would allow government funding only for research on existing, privately derived stem-cell lines. Around a dozen such cell lines are thought to exist, half of them in the United States. The compromise was floated in the press by anonymous White House officials.

President George W. Bush is under mounting pressure from both sides of the stem-cell debate as he moves rapidly towards a decision. And with the administration and Republicans in the Congress openly split on the issue (see *Nature* **411**, 979; 2001), he might welcome a compromise.

But many researchers dismiss the value of access only to existing stem-cell lines. James Thomson of the University of Wisconsin, Madison, who first derived stem cells from human embryos (*Science* **282**, 1145–1147; 1998), calls the idea “a bad compromise” that would “in essence satisfy no one”.

“Research just based on the limited number of cell lines available might be biased,” says Rudolf Jaenisch, a professor of biology at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. “You might miss some important information.” Jaenisch was the senior author on a recent paper (*Science* **293**, 95–97; 2001) showing an undocumented instability in gene expression of mice cloned from embryonic stem cells.

Some researchers concede that federal funding for work on a small number of cell

lines would be better than nothing. Irving Weissman, a biologist at Stanford University in California who has worked on mouse embryonic stem cells, says that, for basic developmental-biology studies, a few lines might suffice. “A lot of good research could go on” if the lines are of excellent quality, he says.

But the political usefulness of the compromise is doubted by some observers. “It doesn’t make any sense, hold any water, or gain the administration anything ethically or politically,” says Tony Mazzaschi, associate vice-president for research at the Association of American Medical Colleges.

Gene Tarne, a spokesman for the Coalition of Americans for Research Ethics, a lobby group opposed to embryonic stem-cell research, calls the compromise objectionable. “The stem-cell lines are derived from destroying embryos, whether that was yesterday or next week,” he says.

Groups representing cell biologists say that different stem-cell lines vary in their ability to grow and differentiate, and that a dozen or so lines would be too few to promise therapies for many diseases. They also point out that several of the existing lines do not grow well in culture, rendering them impractical for research, and that the cells represent only a very narrow range of genetic variation.

Thomson, who produced five of the existing cell lines, notes that his lines were made for experimental purposes. “It’s not clear that they were derived in a way that is appropriate for therapy,” he says.

Another obstacle is that Geron, the California-based biotech firm that funded Thomson’s research, holds an exclusive licence for the use of his cell lines in many applications. ■

Bush plots raid on NIH funds to finance AIDS initiative



Matthew Davis, Washington

When President George W. Bush announced in May that the United States would inject \$200 million into a global fund for fighting AIDS, malaria and tuberculosis, he boasted that the contribution would be “in addition to the billions we spend on research”.

But a month later, with rather less fanfare, addition became subtraction. Bush wrote to Congress last month suggesting that part of his original commitment should be paid for by cutting \$95 million from next year’s funding for the National Institutes of Health (NIH).

According to a memo to Bush from Mitch Daniels, the White House budget director, the transfer would involve a \$25-million tap on the National Institute of Allergy and Infectious Diseases (NIAID), which supports much of the NIH’s AIDS, malaria and tuberculosis research portfolio. The other \$70 million would be removed from the administration’s plan to spend \$306 million on building and improving the NIH’s research facilities.

AIDS activists reacted indignantly to the proposed transfer. “The whole point of this fund is to create a new source of money for battling those three diseases, not to rob Peter to pay Paul,” says Alexis Schuler of the advocacy group AIDS Action.

A spokesman for the NIH says it was “too early to tell” how the White House proposal would affect individual construction projects or the details of the NIAID’s budget.

But sources in both the House of Representatives and the Senate say that Congress is unlikely to accept the idea of money for the new fund being transferred from the NIH. One reason for the likely rejection is that the administration has already sought to use the bulging NIH budget to bankroll other health-related programmes worth \$460 million. ■



Compromised: biologists say that using only existing embryonic stem-cell lines will constrain research.

Royal Society disputes value of carbon sinks

David Adam, London

Britain's Royal Society is pouring cold water on the idea that 'carbon sinks' of forest and farmland could soak up enough carbon dioxide to help industrialized nations meet their long-term emissions targets for reducing greenhouse gases.

Improvements in sink management could see more carbon dioxide sequestered, a report from the society says. But it adds that this would be a short-term, one-off benefit that should not be considered as a long-term alternative to cutting emissions.

The Royal Society's verdict comes as negotiators gather in Bonn, Germany, for the next round of climate-change talks, due to start on 16 July.

A previous round of talks last year in The Hague broke down after European negotiators rejected calls from the United States, Japan, Canada and Australia for a bigger role for carbon sinks.



Sinking feeling: forest management may have a limited effect in reducing the greenhouse-gas mix.

This time round, the Europeans may offer concessions on carbon sinks in a bid to keep the Kyoto Protocol alive in the wake of

US President George W. Bush's decision to abandon it. But the Royal Society report inconveniently challenges the potential usefulness of carbon sinks as a means of addressing the climate-change problem.

"Land carbon sinks may help to reduce greenhouse-gas levels in the atmosphere in the short term, but the amounts of carbon dioxide that can be stored are small compared to emissions from the burning of fossil fuels," says David Read, a plant scientist at the University of Sheffield, who chaired the working group that prepared the report.

Terrestrial vegetation and soil currently absorb about 40% of global carbon-dioxide emissions from human activities. The report agrees that changes to agricultural and forestry practices, such as planting fast-growing 'biofuel' crops and slowing deforestation, could increase this capacity. But it argues that carbon sinks are not a long-term solution because they will quickly become saturated and unable to soak up any more carbon dioxide.

"Managed land sinks could potentially meet 25% of the reductions in carbon dioxide projected to be required globally by 2050," the report says. "However, this would require considerable political will and there is little potential for increasing the land carbon sink thereafter."

The report adds that it is difficult to assess even the short-term usefulness of sinks because of uncertainties in monitoring techniques and suggestions that some sinks could even switch to releasing their stored carbon dioxide in future.

Furthermore, the report warns that some practices designed to boost the capacity of carbon sinks, such as large-scale use of nitrogen-based fertilizers, could actually increase climate change by releasing other greenhouse gases, including methane and nitrous oxide.

Battle to save beleaguered beluga

Sally Goodman, Paris

Police, researchers and fishery managers are to step up their battle to rescue the sturgeons of the Caspian Sea, whose famous eggs account for 90% of the world's caviar.

The illegal harvest of the fish and their valuable eggs — which is thought to exceed the legal catch by an order of magnitude — is depleting supplies of the delicacy.

But conservationists say that the smart use of DNA markers to locate the precise

origins of illegally traded caviar could improve policing and allow the sturgeons, including the beluga (*Huso huso*), to survive.

Last month Russia, Azerbaijan and Kazakhstan — three of the five Caspian Sea states — agreed on a new 12-month action plan to save the sturgeons, at a Paris meeting of the Convention on International Trade in Endangered Species (CITES). Turkmenistan is expected to adhere to the agreement, and CITES declared the fifth state, Iran, exempt because of the quality of its existing sturgeon-management programme.

The states agreed to collaborate with CITES and Interpol, the international network of police. Scientists in the region plan to develop a comprehensive set of the DNA markers that locate the precise source of caviar. The states also pledged to negotiate details of a sustainability programme.

The agreement calls for the cancellation of the autumn 2001 caviar harvest, with export only permitted of fish eggs left over from the spring, on pain of a total caviar-export ban next year. But observers question whether this will halt illegal harvesting.

Sabri Zain, an official at the World Wildlife Fund, warns that any ban on legal exports would weaken pressure on governments to clamp down on illegal fishing and trading. Zain wants to see a labelling system to identify legal caviar.

Scientists and fishery experts meet this week at Oshkosh, Wisconsin, to discuss ways of securing the fish's future.



Nowhere to hide: an illegal trade in caviar is threatening sturgeons in the Caspian Sea.

♦ <http://www.cites.org>

♦ http://www.royalsoc.ac.uk/policy/csink_ann.pdf

UN backs transgenic crops for poorer nations

Mark Schroppe

Rich countries' doubts about genetically modified (GM) crops are damaging to poor countries that urgently need the crops, claims the United Nations' annual report on the status of humankind.

The benefits of GM crops to developing countries are likely to outweigh their risks if their use is properly controlled, says the United Nations (UN) Human Development Report 2001, due out this week. The UN Development Programme has released such reports annually since 1990, but this year's, *Making New Technologies Work for Human Development*, is the first to concentrate on the value of science and technology.

It states that genetic modification and other emerging technologies should be more widely applied to alleviate poverty and malnutrition in poor countries.

The report is not a blanket endorsement of transgenic crops, however. Rather, it recommends that developed countries consider their expanded use on a case-by-case basis. And it says that the risks of GM crops would be best managed if there was more interaction between rich and poor nations, and also if developing countries with experience of GM crops — particularly China — were to share their information more widely.

Omar Noman, deputy director of the office responsible for the report, says the debate on GM crops has been distorted by "scare stories" from environmentalists. The

report says the debate has largely ignored their potential to transform the agriculture of poor countries. But it also notes that the corporations selling the crops have downplayed the difficulties these countries may face in properly monitoring how they are used.

The report highlights the need for rapid development of transgenic crops such as drought- and virus-resistant varieties of the sub-Saharan staple crops sorghum and cassava. But it recommends the mandatory labelling of GM foods so that consumers and nations can make informed decisions. This has been steadfastly opposed by the corporations that own the technology.

The UN's decision to focus this year's

report on technology reflects a growing consensus among experts that economic and human development are underpinned not only by basic needs such as adequate health care and clean water, but also by access to science and technology.

Elsewhere the report recommends eliminating costly government monopolies in telecommunications as well as increasing resources for non-primary education. It also backs the development of "appropriate technologies" such as low-literacy touch-screen computers and low-maintenance fuel cells. And it advocates more efforts to develop vaccines for malaria and AIDS and for less-publicized scourges such as river blindness. ■



Lending a hand: could biotech prevent damage to sorghum caused by the pest *Eldana saccharina*?

GALLO IMAGES/CORBIS

Arctic university gives collaboration pole position

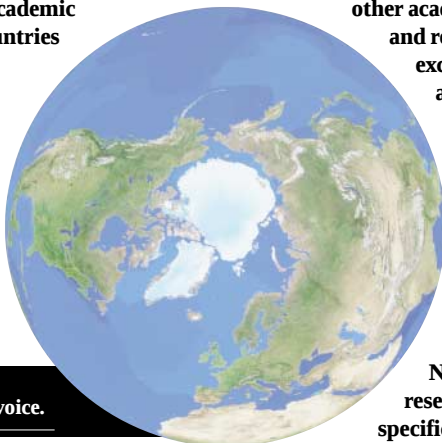
Alison Abbott

A no-walls university in the Arctic region sounds like a pretty chilly prospect. But the newly created University of the Arctic plans to warm up interest in topics of common concern across the region.

The project, which was officially launched last month, is a network of 20 universities and other academic institutions from 10 countries surrounding the North Pole, including Russia, the United States, Canada and the Scandinavian countries.

The fledgling university will offer a degree in circumpolar studies, a multi-disciplinary course that

Opinion pole: the project gives local populations a voice.



covers both natural and social sciences. And it offers several exchange programmes, such as North2north, which is based in Finland, and funds the exchange of students and staff between institutions.

It also includes a research network called the Northern Research Forum (NRF), which brings together young researchers with other academics, policy-makers and resource managers to exchange information and define research issues that need to be addressed.

Jón Haukur Ingimundarson, an anthropologist from the Stefansson Arctic Institute in Akureyri, Iceland, who helps to run the NRF, says that arctic researchers have many specific problems in

common, such as the arctic environment. But researchers in different arctic countries diverge widely in the resources that they have available.

"Icelanders have money, good infrastructure and zero unemployment, for example," he says, "but they need more people. Russians have many highly educated people who remain unemployed and a very poor infrastructure."

The NRF will try to increase co-operation between funding sources and researchers, Ingimundarson adds.

Cynthia Dickson of the Arctic Athabaskan Council, which represents around 30,000 people in Arctic North America, says that being part of the University of the Arctic is a lifeline for the population.

"This university allows us to take part in issues of globalization that affect us directly, like climate change and organic pollution," Dickson says. ■

► <http://www.urova.fi/home/uarctic>

TOM VAN SANT, GEOSPHERE PROJECT/PLANETARY VISIONS/SPL

Nigeria lays plans to launch into the space age

Lagos The Nigerian government has announced plans for a US\$94-million satellite programme.

The initial launches will focus on satellites to improve Nigeria's notoriously poor communications. Subsequent ones will be used for applications in weather prediction and remote sensing aimed at assisting the country's agricultural, forestry and oil-exploration industries.

The government has committed \$27 million to the project in its first year, with a further \$22 million provided annually over the next three years. But some outside observers have criticized the exercise as inappropriate for a country with an external debt of around \$30 billion.

Xenova to have another crack at cocaine vaccine

London Trials of the first-ever vaccine candidate for treating cocaine addiction are set to resume in the United States.

The vaccine, known as TA-CD, stimulates the body's immune system to make antibodies that can attach themselves to cocaine molecules, preventing them from crossing into the brain.

Testing of the vaccine was halted by the US Food and Drug Administration (FDA) in August last year after a related product was shown to cause eye irritation in rabbits. But the FDA has now given permission for the trials to resume, after extensive testing in primates revealed no safety problems.

Xenova, the UK-based pharmaceutical company that is testing the vaccine, also reported that in an earlier trial of nine addicts, cocaine-specific antibodies persisted

throughout the 12 weeks of the study. Participants in the study who relapsed and took cocaine reported that the drug's euphoric effects were reduced.

Biotech firm asks for trials to be kept secret

Sydney US life-sciences company Monsanto has asked the Australian government not to reveal the locations of most of its field trials of genetically modified crops.

The government had insisted that the locations of all trials should be made public by 6 July. But, just hours before the deadline, Monsanto applied to keep the location of 80 of its 85 trials secret. Other participants, including UK company GlaxoSmithKline and two Australian universities, also asked for some of their locations to be withheld, taking the total number of such requests to 111.

Aventis CropScience, a UK-based subsidiary of the French biotechnology company Aventis, agreed to specify the location of almost all of its 283 trials.

Journals open electronic door to poor countries

London Six of the world's largest medical and scientific publishing companies have announced plans to give libraries in almost 100 poor countries free or heavily discounted electronic access to their journals.

The scheme, coordinated by the World Health Organization, will involve nearly 1,000 journals. Access will be made available to medical schools and research institutes in developing countries from January 2002 for at least three years.

Most journals are priced uniformly worldwide. But with annual subscription rates ranging from hundreds to thousands of dollars, many publications are too expensive for research institutions in poor countries.

Participating publishers are Blackwell, Elsevier Science, Harcourt, Kluwer, Springer and John Wiley.

Sect to sue US for right to clone humans

Washington A director of US company Clonaid is to sue the government for the right to continue her research into cloning human beings. Brigitte Boisselier plans to file the suit before Congress votes on pending anti-cloning legislation.

Clonaid was ordered not to try to clone a human being in April after the Food and Drug Administration inspected the company's laboratories. Experiments in animals show that surrogate mothers of clones frequently miscarry and newborns often suffer severe respiratory problems.

Boisselier is a member of a sect called the



Dream on: Boisselier hopes to achieve eternity through cloning.

Raelians. The sect's leader testified before Congress in March that all life on Earth was created by alien scientists and that cloning was the key to the future (see *Nature* **410**, 617; 2001). Boisselier, who says "I believe that one day we will reach eternal life through this method," claims the

company has nearly finished setting up a second lab in another, unnamed, country, where she does not expect legal objections. "If I have to move on, I will," she says.

Recycled nuclear fuel plans leave Japan cold

Tokyo Japan's attempts to use recycled fuel in its nuclear power plants suffered another setback last week, when the Ministry of Economy, Trade and Industry released a report questioning the project's economics.

Japan is planning to burn mixed-oxide fuel — made by extracting plutonium from spent uranium fuel and mixing it with unused uranium — in a third of the country's 51 light-water-cooled nuclear power stations by 2010. But the ministry's report finds that the fuel is more expensive to use than conventional uranium.

The plans were also criticized last week by shareholders of the Tokyo Electric Power Company, which backs use of the recycled fuel. The programme hit trouble earlier this year when residents living near Kashiwazaki-Kariwa power plant voted against use of the fuel (see *Nature* **411**, 729; 2001).

Europe puts stem-cell patents on hold

Munich Applications for European patents on methods for working with embryonic stem cells are to be put on hold until the end of the year, the European Patent Office announced last week.

The office, based in Munich, Germany, has "less than 100" applications pending for stem-cell techniques, says Christian Guggere, its director of biotechnology.

It will wait until the European Group on Ethics in Science and New Technologies, an advisory body to the European Commission, issues an opinion on the subject.

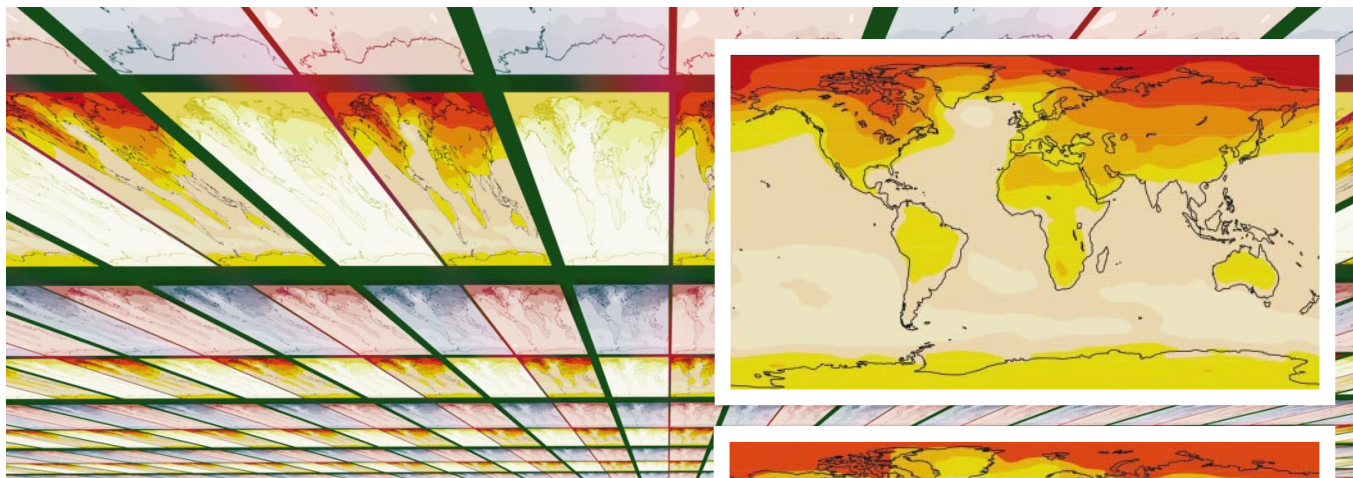
The ethics group last year rejected the creation of human embryos for use in stem-cell research funded by the European Union (see *Nature* **408**, 277; 2000). A follow-up view on patenting issues related to stem-cell research is currently being prepared and will be published by the end of this year.



Snow problem: addicts using a new vaccine find that it reduces cocaine's stimulating effects.

Consensus science, or consensus politics?

To some, the Intergovernmental Panel on Climate Change represents the pinnacle of scientific collaboration. To others, it is a victory for politics over science. Mark Schrope talks to the experts debating our planet's future.



The report on climate-change science finalized in January by Working Group I of the Intergovernmental Panel on Climate Change (IPCC) is no normal review paper. It took 122 lead authors to marshal submissions from 515 other contributors. Twenty-one review editors then took up the baton, ensuring the incorporation of changes suggested by some 700 reviewers. The resulting 881-page document took three years to produce.

Rarely does a single scientific document have such far-reaching consequences for international politics. The report — together with two companions produced by the IPCC's other working groups, dealing with the consequences of climate change and strategies for mitigating them — will be a key reference for delegates who meet in Bonn next week in an attempt to revive the Kyoto Protocol on limiting greenhouse-gas emissions.

Most experts view the IPCC's reports as a huge success story — the first serious attempt to reach a global consensus on a complex scientific issue. But others claim that the involvement of government officials in writing the vital summaries for the reports undermines normal scientific peer-review procedures. Some critics even allege that climate researchers have themselves skewed the reports by expressing their own environmentalist views.

Whatever the truth, such criticisms cannot be ignored. The IPCC aims to provide information to policy-makers without endorsing specific policies. As such, it can only work if it is widely perceived to represent a highly credible and unbiased consensus.

Feeding in the data: the IPCC's reports on the likely extent and effects of global warming are the culmination of gathering, sifting and integrating huge amounts of information.

This year's series of reports was the IPCC's third comprehensive assessment of climate change since it was established in 1988. The arduous process of producing an assessment begins with the assembly of a team of authors for each working group. Ensuring wide international participation is important, because climate change means different things to different countries. A small-island state that risks flooding if sea levels rise may not see eye-to-eye with an oil-exporting nation, for instance. So the IPCC Bureau, the panel's governing body of 30 leading climate experts, invites every one of almost 200 eligible countries to nominate individuals for the working groups. The bureau then makes the final choice so that the lead authors reflect an appropriate international selection while having strong scientific credentials.

Data detectives

The authors then start combing the literature for important papers, and calling on other scientists to submit research that is awaiting peer review or publication. They face the Herculean task of condensing the hundreds of individual submissions into a single report — the inclusion of important

new findings continues almost right up to the final deadline.

The working groups' draft reports are reviewed by a second team of experts appointed by the bureau. Individual countries also have a chance to comment — the first, but not the most controversial, opportunity for politicians to influence the report. For the first and second IPCC assessments, the authors were left to incorporate these revisions. But after complaints that some suggestions were not adequately taken on board, whereas other changes were not reviewed by the original authors, this time a third body of scientific experts — the review editors — was appointed by the bureau to oversee the revision process.

Given the reports' size — the third full assessment covers about 2,500 pages — few people ever read a full working-group report, much less an entire assessment. Instead, most rely on the 'Summary for Policymakers' (SPM) that accompanies each working-group report. For many critics of the IPCC process, the SPMs, and the way in which they are interpreted in the media, are the key problem.

Drafting an SPM is extremely difficult. When circulated for comment, the draft

SPM for Working Group I's third assessment attracted roughly 20 words of comment for every word of the original. Once these comments have been incorporated, the fun really begins. Each working group's SPM has to be approved by a special plenary meeting. This time, roughly 50 authors attended each plenary, along with representatives from non-governmental organizations (NGOs) — but both the scientists and NGOs are there only to advise. The final word rests with 400-odd delegates from the participating countries, who may or may not have a strong background in a relevant scientific discipline.

A literary circus

Every word has to be agreed on unanimously before it enters the SPM. With discussions translated simultaneously into five different languages, the approval of Working Group I's summary for the third assessment — originally just seven pages long — took four days. By the time the delegates to the plenary, held in Shanghai in January, had finished, the document had more than doubled in length.

"You've got an almost circus-like atmosphere," says Thomas Karl, director of the National Climatic Data Center in Asheville, North Carolina, and a Working Group I lead author who was present in Shanghai. "It's hard...very hard."

As government representatives, delegates may arrive with goals for what the summary should say that are based on policy objectives rather than science. "They are trying to perhaps weight certain things based on their national interests," says Cynthia Rosenzweig, a Working Group II lead author who heads the Climate Impacts Group at NASA's Goddard Institute for Space Studies in New York.

Kevin Trenberth, an atmospheric scientist at the National Center for Atmospheric Research in Boulder, Colorado, and a lead author for Working Group I, says that the wrangling in Shanghai began with the first paragraph. Saudi Arabian delegates, whom

some allege are keen to play down the threat of global warming to protect their country's oil exports, objected to a sentence that read: "Many hundreds of scientists contributed to its preparation and review." The Saudis felt this implied that all of these scientists endorsed the report in every respect. After some verbal wrestling, the line was altered to read: "Many hundreds of scientists from many countries participated in its preparations and review." In the days of discussion that followed, says Trenberth, the process was repeated over and over again.

The most difficult negotiations, according to Trenberth, related to the connection between human activity and climate change. One contentious line began as: "Despite these uncertainties, it is likely that increasing concentrations of anthropogenic greenhouse gases have combined substantially to the observed warming over the last 50 years." By the end of the plenary, it had become: "In the light of new evidence and taking into account the remaining uncertainties, most of the observed warming over the last 50 years is likely to have been due to the increase in greenhouse gas concentrations."

To the untutored eye, these slight differences in wording may seem of little significance. And to Rosenzweig, the fact that such protracted negotiations result in relatively limited changes is a testament to the quality of the IPCC scientists' work. By defining the scientific consensus between narrow limits, she argues, there is little room for bias in the summary. Rosenzweig says the biggest problem was coping with scientists pushing to have their own results included.

Indeed, despite the unusual nature of a process that involves political appointees agreeing on how a scientific study should be summarized, most of the IPCC's authors are satisfied with the way it works. But many say that the SPMs should be marked to indicate that they are not solely the work of scientists. Robert Watson, director for the environment at the World Bank and chair of the IPCC Bureau, says he would not object to some words of explanation being included.

Clarity or bias?

John Houghton of Britain's Hadley Centre for Climate Prediction and Research in Bracknell, west of London, and co-chair of Working Group I, believes the plenary meetings actually improve the final SPMs. "It's more clear and more relevant," he says. "You might think it would make it worse, but it doesn't."

But speak to critics outside the IPCC's fold and you hear a different story. Fred Singer is a long-standing sceptic of the threat posed by global warming, and president of the Science and Environmental Policy Project, a pressure group in Arlington, Virginia. Singer thinks the main reports are sound, but he argues that the SPMs fail to



Up for discussion: John Houghton believes the plenary sessions improve the final report.

adequately acknowledge the uncertainties in climate-change science. According to Singer, the process that gives rise to the SPMs plays down uncertainties so as to force governments to take climate change seriously. "It is selective in the facts that it uses from the report," says Singer. "It slants things. It puts a spin on things. It starts out with a given conclusion and then selects those facts which support that conclusion."

Most climate researchers argue these charges should be laid at the door of Singer's group, not the IPCC. Thomas Stocker, a climate modeller at the University of Bern in Switzerland and one of Working Group I's lead authors, takes serious issue with Singer's broadside. "That's completely unfounded," he says. "If you read our SPM you will find plenty of sentences that explicitly state where there are uncertainties." He views the agreement between the full report and its précis in the SPM as impressive.

Reflected glory

Exactly how accurately the SPMs reflect the underlying reports was investigated last month by the US National Academy of Sciences (NAS) as part of a request from President George W. Bush to examine climate-change research (see *Nature* **411**, 725; 2001). The NAS report focused on Working Group I — the most contentious, as the other two working groups depend in large part on its conclusions. Like the working group itself, the NAS report provided ammunition for both supporters and critics of the IPCC. Working Group I's SPM, concluded the NAS, is "consistent with the main body of the report". But the NAS report agreed that the SPM had failed to explain adequately the caveats on which some of the uncertainties it referred to were based.

"We found some understatements of uncertainties, but generally changes made from the technical chapters to the SPM didn't affect the impact of the statements very much, which was very impressive," says Ralph Cicerone, an atmospheric scientist and chancellor of the University of California, Irvine, who chaired the NAS panel. Cicerone adds



No doubt: Thomas Stocker feels that the IPCC's reports make clear any uncertainties.

▶ that his panel conducted a quick survey of IPCC authors in the United States and found that those who responded said unanimously that the SPM accurately represented what they wrote in the main text.

But Richard Lindzen, a meteorologist at the Massachusetts Institute of Technology and an author of the NAS report, does not see it the same way. "Within the confines of professional courtesy," Lindzen wrote in an opinion piece in *The New York Times*, "the panel essentially concluded that the IPCC's Summary for Policymakers does not provide suitable guidance for the US government." Lindzen has long argued that evidence for climate change is too shaky to justify the costly strategies mooted to tackle it.

Some critics of the IPCC believe that removing politicians from the process could be one way of ending the arguments. "Let the scientists tell the world what the scientists said," says Robert Balling, director of the Office of Climatology at Arizona State

It might look like a circus at times, but a global response to climate change would probably be impossible without the IPCC.

University in Tempe. Although Balling contributed material to the Working Group II report, he is a prominent global-warming sceptic and a vocal critic of the IPCC.

But most participants in the IPCC process believe that the presence of national delegates is crucial. "If you didn't have that then the report wouldn't be so important. It's as simple as that," says Michael Grubb, an energy economist at Imperial College London and a Working Group III lead author. He argues that involving government officials in the IPCC process forces politicians to take a close look at the report and the underlying science, rather than simply putting it on the shelf.

Michael Oppenheimer, chief scientist with the New York-based NGO Environmental Defense, and a lead author for Working Group I, agrees. "If the price paid to get there is some discomfort among some of the players and criticism because it's not purely a product of scientists, I think that's a price worth paying," he says.

Panel beating

In the past, revisions to the working groups' full reports subsequent to SPM approval have caused problems. When substantial additions are made to the draft of the SPM, authors must sometimes supplement the main report to ensure that a given point is adequately covered. After the IPCC's second assessment was published, Singer's Science and Environmental Policy Project alleged that this process had proceeded largely unchecked, with some authors adding material to the main report without giving others the chance to comment. To deflect such criticism of the third assessment, every change to the main report was logged, whether it resulted from the formal review process or the plenary sessions.

Some critics allege that problems of bias lie not with the SPMs or the IPCC process, but with the climate research community as a whole. They argue that many climate scientists hold environmentalist views, and so tend to stress the importance of research that paints the most worrying picture of climate change in order to spur politicians into action.

The Global Climate Coalition (GCC), based in Washington, which styles itself as a "voice for business in the climate debate", draws attention to remarks made by Stephen Schneider, a climate-change researcher at

Stanford University in California and a lead author of the Working Group II report.

In a 1989 *Discover* magazine article, Schneider discussed the dilemma facing scientists who wanted to draw attention to climate change while remaining true to current scientific knowledge of the subject. "We need to capture the public's imagination," he noted. "That entails getting loads of media coverage, so we have to offer up scary scenarios, make simplified and dramatic statements and make little mention of any doubts we might have." Although Schneider went on to say that he hoped climatologists could be both effective and honest, the remarks were seized on by bodies such as the GCC as evidence that scientists were exaggerating the consequences of climate change.

Tempered view

John Christy, an atmospheric scientist at the University of Alabama in Huntsville, and a lead author for Working Group I, believes that researchers who are new to climatology, many of whom have been attracted to the field because of its 'green' associations, are more likely to back extreme climate-change predictions. But he argues that they are largely countered by more experienced researchers who tend to make more conservative judgements because they place less stock in the long-range climate forecasts.

The fact that the IPCC's consensus is backed by Christy, whose views on climate change have on occasion provided ammunition for global-warming sceptics, provides one indication that — despite its critics — the organization is working effectively. But where the IPCC goes from here is still being debated. A fourth assessment will almost certainly take place, but Watson says it might focus on new information rather than attempt another full review.

Some researchers, meanwhile, suggest that the IPCC model should be applied to other thorny issues. Christy, for example, would like to see a similar effort devoted to the environmental problems threatening developing nations, such as deforestation and the lack of suitable fresh water. And the UK House of Commons' Science and Technology Committee recently suggested that the IPCC model could be applied to ocean pollution and genetic modification. "It's the only long-term successful example of how a complex scientific issue can be brought to the decision makers," says Stocker.

Whether or not the model is applied elsewhere, the fact that IPCC reports are accepted as the scientific guide for the Kyoto Protocol negotiations is, for many involved, proof enough that the process is working as well as can be hoped. It might look like a circus at times, but a global response to climate change would probably be impossible without it. ■

Mark Schroppe is a freelance writer in Melbourne, Florida.

▶ <http://www.ipcc.ch>

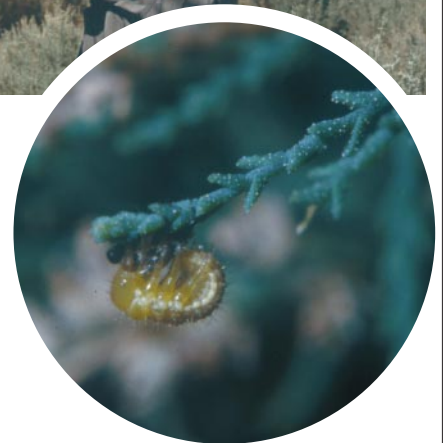
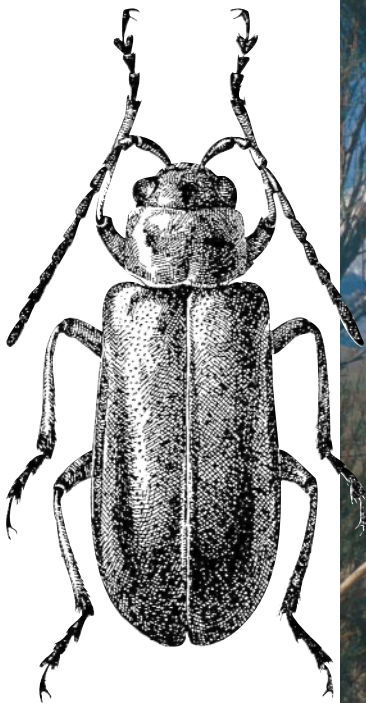


Under review: could an IPCC-style approach help assessments of lack of water in developing countries, deforestation or genetic modification?

JOE RAEDLE/NEWSMAKERS

J. C. VINCENT/STILL PICTURES

NICK COBBING/STILL PICTURES



Alien versus predator

Can invasive species be controlled by introducing their natural enemies? The idea has a chequered history. But as safety testing improves, it is now gaining currency. Jonathan Knight reports.

Throughout the world, exotic aliens are wreaking havoc. From Japanese knotweed in the wetlands of Europe to Russian wheat aphids in the plains of North America, invasive species spread to new habitats by human activities are causing serious harm to ecosystems and agriculture. One estimate puts the annual cost at \$137 billion in the United States alone¹.

Invading species often do well in their new environments because they lack the predators that kept them in check back home. So, for more than a century, people have tried dealing with problem exotics by importing more exotics to feed on them.

The results have been mixed — about one-third of species introduced to control exotic weeds have been judged a success². The Australian experience illustrates the extremes. In a notable success, the cactus moth (*Cactoblastis cactorum*) was imported from Argentina in 1925 to deal with an infestation of prickly pear (*Opuntia* spp.). The moth reclaimed some 6.5 million hectares of rangeland for agriculture in just a few years without harming a single native Australian plant. But this must be set against infamous ecological disasters. In one example, cane toads (*Bufo marinus*) imported to northern

Queensland in 1935 showed little appetite for the sugar-cane beetles they were meant to control. Instead, they have munched their way through much of the region's unique wildlife.

Although critics still worry about the potential for things to go wrong, recent improvements in safety testing before release have begun to ease fears about biocontrol getting out of control. And in May, scientists with the Agricultural Research Service (ARS), the research arm of the US Department of Agriculture (USDA), released what is thought to be the most carefully tested and monitored biocontrol agent ever.

The ARS is deploying Chinese leaf beetle (*Diorhabda elongata*) against saltcedar trees (*Tamarix* spp.). These trees, native to Europe and the Middle East, now infest river banks throughout the western United States, where they are displacing native willow and cottonwood. Their deep roots draw up salt, raising topsoil salinity and killing undergrowth.

In theory, biological control is an elegant solution to an intractable problem. It is usually much more specific to the target than pesticides or herbicides, and if the control organism takes hold, it is self-perpetuating. But introduced organisms cannot be with-

Beetlemania:

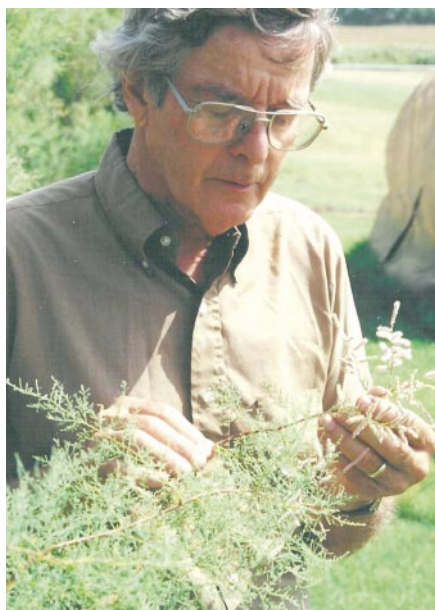
Tom Dudley (top, right) releases *Diorhabda* (top left) in California in an effort to combat saltcedar. The insect's larvae (inset) seem to feed almost exclusively on the trees.

drawn. Even if an exotic biocontrol agent has no damaging consequences at the initial site of release, its spread can create problems. After the cactus moth was introduced to Nevis in the Caribbean in 1957 to control prickly pears, it spread from island to island. Now it threatens rare native cacti in southern Florida. Ultimately, it could reach Mexico, where prickly pear is an economically important crop.

Ground force

Biological control of invasive species is gaining in popularity. Peter McEvoy, an entomologist at Oregon State University in Corvallis, found that the number of agents released to control invasive weeds in his state doubled to nearly 70 in the decade ending in 1997 (ref. 3). The trend is similar around the world⁴.

Traditionally, the prime motivation has been to protect agriculture, which has outweighed concern for native species. The Eurasian weevil (*Rhinocyllus conicus*), for



Low risk: Jack DeLoach believes *Diorhabda's* eating habits will make it safe to release.

instance, was released in the United States and Canada during the late 1960s to control introduced thistles that were overrunning grazing land, even though these invaders were growing beside native thistle species. In 1997, field biologists in Colorado and Nebraska found that the weevil was feeding on the seeds of seven native thistle species in national parks and nature reserves⁵.

But growing sensitivity to the need to protect native species has led to a tightening of the rules, says Tom Dudley, who works on biological control at the University of California, Berkeley. "If the thistle project were proposed right now, there is no way it would be approved," he says.

One-track minds

The key to safety in biological control is target specificity. The best targets are those that have no close relatives in the area they have invaded, and the best control agents are those that eat one thing only.

The saltcedar control programme scores well on both criteria. Prickly pears in Australia made a good target because they had no close relatives, and saltcedar is in a similar position in the western United States. Several species were imported in the nineteenth century as windbreaks and ornamentals. Two of them, *Tamarix parviflora* and *T. ramosissima*, are now causing serious ecological harm. Because they have no close relatives in North America, Jack DeLoach, a biologist at the ARS's station in Temple, Texas, was able to convince a federal regulatory board in the late 1980s that biological control would pose a low risk.

The USDA had started looking for suitable enemies to deploy against *Tamarix* in 1987. Two years later, DeLoach's team identified the *Diorhabda* leaf beetle, in western China, where it had been thwarting local

attempts to grow saltcedar for erosion control by consuming its foliage. Over the past nine years, under conditions of tight quarantine in Temple, DeLoach and his colleagues have tested the beetle on more than 100 plants native to western North America. *Diorhabda* showed little appetite for any of them. "There is no case of anything being tested as carefully as this," says Dudley, who works on the saltcedar team. "We can never be 100% sure of no negative effects, but we are pretty darn sure at this point."

Researchers who have been critical of previous attempts at biological control agree that *Diorhabda* is probably safe to release. "In this one case I'm not too worried," says Daniel Simberloff, a specialist in the biology of invasive species at the University of Tennessee in Knoxville. "They have been very good about testing."

The saltcedar programme is also notable for its long-term follow-up plan. Past attempts at biological control have released different species in rapid succession with little subsequent study, in the hope that at least one would do the job. But in 1999, the ARS began requiring any new weed-control programme to develop a long-term management plan to measure its effectiveness—and, if necessary, to help replant with native species. Only after one agent is deemed inadequate will the next be brought in.

A long-term management plan is particularly relevant in the case of saltcedar. Trees take much longer to kill than herbaceous weeds, as they can survive several years of defoliation before they begin to die back. The saltcedar monitoring programme will also monitor the progress of the southwestern willow flycatcher (*Empidonax traillii extimus*). This endangered bird has taken to nesting in saltcedar in parts of Arizona, and conservationists are worried that killing the trees will leave it with nowhere to breed. The May releases of *Diorhabda* were made at least 200 kilometres from the bird's saltcedar nesting sites, and ARS researchers will be watching how quickly native willow and cottonwood replace any dead saltcedar. If the return of native species is too slow, the USDA and other agencies will consider planting trees to ensure that the birds have a place to nest.

Although ecological safety is now an issue in weed-control programmes, the application of similar standards to the biocontrol of invasive insect species has lagged behind. The USDA's Animal and Plant Health Inspection Service (APHIS) has been liberal in granting release permits, and it is a similar story around the world.

Fewer than 400 invertebrates and fungi have been released for weed control, anywhere in the world³. But some 5,000 organisms have been released worldwide in attempts to control exotic insects, according to a database compiled by CAB International, a non-profit organization based in

Wallingford, Oxfordshire, that promotes the development of sustainable agriculture.

Between 1986 and 1993, for example, APHIS released 29 insect enemies to attack the Russian wheat aphid (*Diuraphis noxia*), a serious problem for Midwestern farmers. None has been particularly effective, but at least one, the seven-spot ladybird (*Coccinella septempunctata*), now appears to be pushing out native ladybird species⁶.

Food for thought

In the United States, at least, rules introduced under the Plant Protection Act of July 2000 now require insect-control researchers to provide APHIS with a full Environmental Assessment before they are permitted to release predatory insects. The new rules have "certainly raised everyone's awareness and sensitivity", says Russell Messing, a biocontrol researcher at the University of Hawaii on Kauai.

If this increased sensitivity leads to insect-control programmes adopting the rigorous methods used in the saltcedar project, the chances of repeating the cane-toad disaster should be greatly reduced. That would be welcome, as the threat posed by exotic species is growing all the time.

"Not doing anything is clearly a very negative process for the environment," says Ray Carruthers, who heads the ARS's Exotic and Invasive Weed Research Unit in Albany, California, which administers the saltcedar project. "We either sit here and watch it happen, or try to develop effective programmes that have minimal environmental risk."

Jonathan Knight writes for *Nature* from San Francisco.

1. Pimentel, D., Lach, L., Zuniga, R. & Morrison, D. *BioScience* **50**, 53–65 (2000).
2. Williamson, M. & Fitter, A. *Ecology* **77**, 1661–1666 (1996).
3. McEvoy, P. B. & Coombs, E. M. *Ecol. Appl.* **9**, 387–401 (1999).
4. Julien, M. H. & Griffiths, M. W. (eds) *Biological Control of Weeds: A World Catalogue of Agents and their Target Weeds* (CABI Publishing, Wallingford, UK, 1998).
5. Louda, S. M., Kendall, D., Connor, J. & Simberloff, D. *Science* **277**, 1088–1090 (1997).
6. Obrycki, J. in *Nontarget Effects of Biological Control* (eds Follett, P. A. & Duan, J. J.) 31–44 (Elsevier, Dordrecht, the Netherlands, 2000).



Toad haul: the cane toad eats almost anything except the beetles it was meant to control.

Seeking, sometimes finding, that elusive chemistry

Despite all the discipline's achievements, opinion is divided as to whether chemistry is getting the recognition it deserves — and needs — in order to keep attracting new talent.

Sir — The Opinion article “A discipline buried by success” (*Nature* **411**, 399; 2001) and News Feature “What’s in a name?” (*Nature* **411**, 408–409; 2001) in the 24 May issue are correct in their analysis of the lack of recognition of chemistry, in and outside the international scientific community. Scientists, policy-makers and the general public should take note of these timely messages.

I would like to add that the lack of recognition for the breadth of modern chemistry in China is hurting chemistry and related fields.

On 12 May 2001, the Nobel laureate Harry Kroto delivered a lecture titled “Science: a round peg in a square world”, at the Great Hall of the People at Beijing, in which he passionately called for both better understanding of the role of basic research and better public understanding of scientific ideas. I was delighted to serve as Professor Kroto’s translator, and accompanied him to a discussion with 50 high-school students at No. 4 High School in Beijing, one of China’s few elite schools.

One student asked why, with biology in the ascendant, she should study chemistry. Part of Kroto’s answer was that understanding and controlling chemistry at the molecular level is the key to the success of molecular biology and molecular electronics. This information was new to these bright young students, who will soon be choosing their careers.

In China, the lack of recognition of the breadth of chemistry is alarming. Biochemistry, for example, has never been a discipline within chemistry. The Chinese Chemical Society (CCS) does not have a biochemistry division, and the chemistry division of the National Natural Science Foundation does not support biochemistry research. The recent hype about state projects on the human genome sequence and related fields (*Nature* **410**, 10–12; 2001) excludes the involvement of chemists. The president of the American Chemical Society told me in Beijing that more than 50% of the society’s members are industrial chemists: in contrast, there is not even a Chinese word for ‘industrial chemists’. China does have ‘chemical engineers’, but they are not covered by CCS membership.

Modern chemistry is about much more than beakers and flasks. The

discoveries of buckminsterfullerene (C_{60}) and carbon nanotubes have reminded us that chemical synthesis can be done with sophisticated machines. The widely used technique of electrospray mass spectrometry in medical screening and biological analysis was developed and perfected in physical chemistry laboratories.

Yet the Chinese science community and China’s educational administrators have failed to recognize many of these facts — which is largely why the country’s undergraduate and graduate chemistry programmes are outdated.

One of the direct consequences is that my laboratory cannot find students with decent training in modern physical chemistry.

Graduate students and postdocs from China have become a sizeable part of the research force in many US and European research institutions, so China’s lack of modern chemistry skills is also a loss to the world at large.

Hong-fei Wang

State Key Laboratory of Molecular Reaction Dynamics (Beijing), Centre for Molecular Sciences, Institute of Chemistry, Chinese Academy of Sciences, No. 2, 1st North Street, Zhongguancun, Haidian District, Beijing 100080, China

Researchers are popular, even if the industry is not

Sir — Your disappointing Opinion article and News Feature (*Nature* **411**, 399 and 408–409; 2001), bemoaning the poor public image of chemists and chemistry, do not refer to a recent survey carried out by Wirthlin Worldwide and sponsored by the American Chemical Society (see <http://www.acs.org/wirthlin.html>). This research indicates that the US public views chemists favourably in many ways, associating them with being visionary, innovative and results-oriented.

Although concerned about the effects of chemicals on their everyday health and safety, respondents also had positive feelings about a range of chemistry’s contributions to everyday life, from agriculture to cleaning products. And although your articles suggest that chemistry achievements such as new pharmaceuticals are “being

appropriated by other disciplines”, more respondents to our survey credited chemists with that achievement than with any other.

Nearly 18,000 international scientists attended the ACS meeting that was featured in your Opinion article, not 1,000, as *Nature* reported.

Denise Graveline

American Chemical Society, 1155 16th Street NW, Washington, DC 20036, USA

We apologize for the inadvertent error in our reporting of the number of scientists at the meeting, which was introduced during editing — Correspondence Editor, *Nature*.

Time to shout about the benefits of chemistry

Sir — In your interesting Opinion article on chemistry (*Nature* **411**, 399; 2001), you comment that in my Perspective article “The Quiet Revolution in Chemistry” (*Chemical and Engineering News* 64–65, 7 August 2000), I stop short of identifying potential applications. This is not so.

In my Perspective I identify how, by achieving one or more of the objectives on my ‘wish list’, chemists could contribute significantly to improving the human condition.

I list three of many possible applications in an ‘imagine’ list: imagine bridges that do not corrode; imagine Rome, Bangkok and Los Angeles with no air pollution, and with tap water that you would enjoy drinking; and imagine learning the entire health profile of a person from a drop of blood.

My point was to highlight some of the grand challenges of fundamental chemistry, which many believe are as exciting and important as similar challenges in our sister fields of biology and physics.

I used the term “quiet” in the sense that the science media are not fully aware of these revolutionary objectives, and *Nature* is to be applauded for helping to make the revolution more noisy.

Stephen J. Lippard

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Singapore makes efforts to sustain biodiversity

Sir — Agoramoorthy and Hsu in Correspondence (*Nature* **410**, 144; 2001) express the fear that, although Singapore is creating a genome project costing S\$62 million (US\$35 million) over five years, it is ignoring some serious environmental issues.

They suggest that the number of future students of ecology and conservation in Singapore may fall, and rightly point out that, as an economic giant, Singapore has the responsibility to care for its natural environment and to help to conserve natural resources in neighbouring 'mega-diversity' countries.

They mention the government's conservation fund and Wildlife Reserves Singapore. These are not, however, the only large and long-standing conservation efforts under way. For example, the National University of Singapore (NUS) has been very active in conservation work and biodiversity research for some 30 years. In 1998, the Raffles Museum for Biodiversity Research was established at the NUS, with a budget of about S\$2.7 million, specifically to promote research on biodiversity and to deal with conservation issues in collaboration with the Singapore National Parks Board. These institutions have hosted hundreds of researchers from many countries who study biodiversity in Southeast Asia.

Between 1994 and 1999, NUS staff and students published some 300 articles, many in the leading scientific journals of their fields (Ng, P. K. L. *National Science Museum Monographs (Tokyo)* **18**, 2–23; 2000). The NUS itself publishes the *Raffles Bulletin of Zoology*, an international, peer-reviewed journal of faunal studies in Southeast Asia. The Raffles Museum for Biodiversity Research has been involved in joint research and training projects with neighbouring countries, notably Brunei, China, Indonesia, Malaysia, the Philippines, Sri Lanka, Taiwan, Thailand and Vietnam. Further, a new public exhibit gallery aimed at promoting public awareness and education in conservation opened on 15 June 2001 (see <http://www.rmb.rnus.edu.sg>).

We have been impressed by the enthusiasm of numerous graduate students dedicated to ecology, biodiversity and conservation research at the NUS. Contrary to the fears of your correspondents, much emphasis has been placed on such issues at the NUS over the past decade and we think that Singapore is well equipped to tackle them.

We hope that the government of Singapore will continue its commitment to

and funding of biodiversity research and conservation in this rich island nation.

Lanna Cheng*, Damir Kovac†

*Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093-0202, USA

†Forschungsinstitut Senckenberg, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

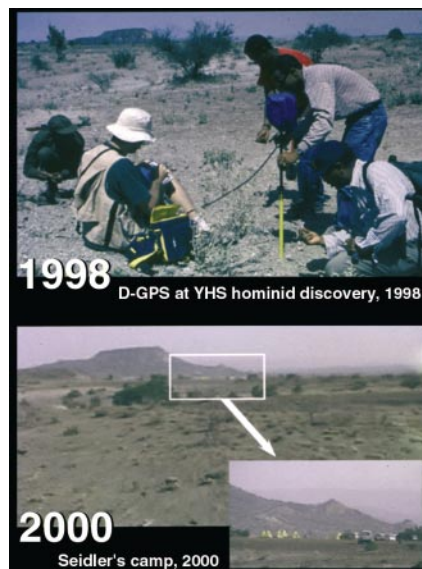
Photos may offer clues over Ethiopian fossil site

Sir — Horst Seidler, in his Correspondence (*Nature* **411**, 15; 2001) responding to the News story "Restrictions delay fossil hunts in Ethiopia" (*Nature* **410**, 728; 2001), states that his use of the Galili research site in Ethiopia is legal. But I, not he, discovered this site, as shown in the accompanying photograph.

The Ethiopian Authority for Research and Conservation of Cultural Heritage, which regulates the use of sites, incorrectly assumed that Galili and the Mulu Basin are different, and has been persuaded to allow Seidler's occupation to continue.

After participating in the discovery of Galili in 1997, I worked there each year under official permit. Then, after my first publication from this site (*Am. J. Phys. Anthropol.* **111**, Suppl. 30, 170; 2000), Seidler arrived there, camping less than 200 metres from the place where I had found hominid fossils two years earlier. Now he claims, incorrectly, that my permit was for a different area, although the photographs below clearly show his camp in the same area as my earlier site.

Seidler notes that after arriving at my site, he offered to let me join his team. This



Digging in: Haile-Selassie's Galili camp, top, and Seidler's camp, bottom and inset.

offer was completely inappropriate as the site was mine in the first place. It would have been far better for Professor Seidler to have withdrawn from Galili when I appealed to Ethiopian regulators about his team's arrival there. Instead, he first apologized for what happened, now claims that my permit was for a different area, and continues to occupy the site.

Yohannes Haile-Selassie

Laboratory for Human Evolutionary Studies, Department of Integrative Biology, University of California, Berkeley, California 94720, USA

Keeping Mendel in mind

Sir — Science is a function of cultural and social imprinting, the impression left by the environment in which a researcher lives and works. Johann Gregor Mendel lived in difficult circumstances, with his work neglected during his lifetime, and disregarded under the influence of Trofim Lysenko. Now, as your News story "Museum suffers spiritual cramps over Mendel's work" (*Nature* **410**, 6; 2001) makes clear, the Mendel museum in Brno, the Mendelianum, is under threat on rather flimsy religious grounds.

After a recent international human genetics meeting in Vienna, I visited the the Mendelianum and the abbey where Mendel designed and performed the experiments forming the cornerstone of genetics. The museum is managed by cordial, enthusiastic people, but it is small and has limited resources compared with similar institutions in Eastern Europe, for example the museum dedicated to Ignaz Semmelweis in Budapest. Few of my colleagues at the Vienna meeting were interested in visiting the cradle of genetics, and there is an air of indifference in this lovely city towards such an illustrious citizen and his memorial.

Given the relevance and impact of genetics, most recently with the decoding of the human genome sequence, the scientific community, especially in Eastern Europe, must support the Mendelianum. A fund could be established, genetics meetings could be held in Brno, and promotional material could be distributed to research centres, museums and other institutions across the world. Hence, future generations will become culturally 'imprinted', not with the dogmatic ignorance of Lysenko's heirs, but with the work of this universal figure to whom we, as a rational species, owe so much.

Fabio Salamanca

Unit of Medical Research in Human Genetics, National Medical Center, Mexican Institute of Social Security, Apartado postal 12-951, Mexico DF 03020, Mexico

Eccentric origins of creativity

Did the genes underlying schizophrenia drive human evolution?

**The Madness of Adam and Eve:
How Schizophrenia Shaped
Humanity**

by David Horrobin

Bantam: 2001. 275 pp, £18.99

Daniel Nettle

David Horrobin's preface promises us an intellectual Odyssey, and inasmuch as the range of the Odyssey was great, and its feats bold, his promise is fulfilled. This book contains more than just one bold argument — it has actually three connected but quite distinct major theses in one slim volume.

The first thesis starts from the undeniable truth that schizophrenia is an evolutionary puzzle. It is substantially genetic, moderately common in all human populations, and known to be associated with reduced reproductive success. How, then, do the genes underlying it persist at such frequency?

Horrobin's answer is that the traits that underlie schizophrenia are also found in many individuals who are not ill and who are among the most creative and successful members of society. Thus, the genes are maintained by positive selection on those who avoid frank illness, balanced by negative selection on those who do become ill.

Evidence for the co-occurrence of psychosis and creativity in families has been amassing for some time. Many links in the chain of causation still need to be filled in more solidly, though. For example, although theories involving the evolution of creativity rest on the notion that creative people enjoy increased reproductive success, this has never been demonstrated empirically for any modern or traditional society. In traditional societies, much of the variation in reproductive success can be explained simply by wealth, and neither artists, original though they are, nor schizotypal groups — those with behavioural eccentricities similar to but less severe than seen in true schizophrenia — tend to be high achievers in narrow financial terms.

The book's second thesis is that the neurobiological basis of schizophrenia concerns the metabolism of fatty acids. This radical claim has hovered at the margin of psychiatric research for some years. Horrobin, its chief evangelist, argues eloquently that the time has come to give it a more central position. The model can in fact be reconciled with more conventional explanations involving the chemical transmitters between neural cells, or altered neuroanatomy, as fatty acids are involved in both neurotransmission and the growth of neural circuitry. Most significantly, the fatty-acid model



offers at least a possibility for developing new treatments in an area where medical progress has been stubbornly slow.

Horrobin's third thesis is more troublesome. He subscribes to the 'Big Bang' theory of human evolution, which tries to identify a key change at the origin of anatomically modern humans that gave rise to the full flowering of art, culture, language, technology and global conquest in a single step. Horrobin's version of the great leap forward is a change in fatty-acid production that radically enhanced neural connectivity within the brain. These very changes, Horrobin argues, are those associated with schizophrenia. He writes: "Without the genes which in combination cause schizophrenia we would be like Neanderthals or *Homo erectus* — large-brained, clever, but lacking that lust for change and creative spark that have so dramatically distinguished our species from our immediate predecessors".

This idea is problematic. Language, symbolism, culture and technological change are universal properties of all humans, whereas the genes that produce schizophrenia are restricted to a small minority of the population. Schizotypy cannot therefore be a central part of what distinguishes us from other hominids. Indeed, there is a deeper problem, for if schizotypal genes have an overwhelming selective advantage, as Horrobin supposes, then they should become universal in the population. But this has not happened. The real evolutionary question is why the genes relating to psychiatric disorder are neither vanishingly rare nor universal, but occur at a troublesome middling frequency, too common to be explained by genetic drift



In the eye of the beholder: creations born of schizophrenic minds.

and too rare to be unequivocally adaptive.

Although he expounds his thesis elegantly, Horrobin is occasionally unclear about the logic of his evolutionary explanation. He dismisses previous accounts of human evolution, which promote social complexity or sexual selection as primary factors, in favour of his hypothesized change in phospholipid metabolism. However, the hypotheses he dismisses concern the ultimate causes of evolutionary change — selection pressures — whereas his model is one of proximate causation. An account of selection pressures is still required to explain why it is so good to be creative, and here we must fall back on such factors as sociality and sexual selection once again.

This is a popular work rather than a monograph, which has both advantages and disadvantages. On the positive side, the book is vivid, sweeping and readable, although sometimes paragraphs repeat themselves almost identically within a few pages of one another. On the negative side, the level of scientific attribution is very poor. There is an extensive bibliography, but it does not follow the structure of the book. As a result, it is impossible to trace the references to the studies alluded to in the text. Some results are not attributed at all, which is a shame, because the writing clearly comes from a good knowledge of the literature. Overall, the book is highly stimulating, but perhaps overreaches itself when it tries to turn an interesting and plausible argument about psychosis into an explanation of how we as a species evolved. ■

Daniel Nettle is at 10 Bears Hedge, Oxford OX4 4JJ, UK.

An unsung hero put on the map

The Map That Changed The World: William Smith and the Birth of Modern Geology

by Simon Winchester
 Viking: 2001. 338 pp. £12.99

Douglas Palmer

In 1797, William Smith, a surveyor and the orphaned son of an Oxfordshire blacksmith, produced the first list of the rock strata in the west of England; 1815 saw the publication of his geological map of England, the first of any country on a scale of five miles to the inch. Four years later Smith was in King's Bench Prison in Southwark for debt. And in 1831, nine years after his release, the Geological Society of London awarded Smith the first Wollaston Medal and he received a pension of £100 a year from King William IV.

But who today, outside the world of professional geology, has heard of this quintessentially English technical genius, a 'doer' rather than an intellectual? One reason is the lack of a modern biography of Smith. Simon Winchester now makes an honourable attempt to tell this remarkable story of scientific achievement against enormous odds: how Smith's great map was plagiarized, his financial failure and his belated recognition. The book will help to bring Smith the wider attention he deserves, as unsung heroes are in short supply. But there are considerable difficulties in writing a popular biography of a man such as Smith, difficulties that I suspect have deterred previous would-be authors.

There is little documentation on Smith's early years. Most of his professional life was spent trudging around the country surveying land, obsessively looking at rocks and making lists of the strata and their fossils—hardly the riveting stuff of popular biography. Smith did not write easily, although he did write a good deal, and his achievements are largely encapsulated in his beautiful and innovative geological maps. These were a huge technical achievement, but their niceties can be difficult for the uninitiated to appreciate. The problem for the biographer is to explain all this, to put it in a wider historical context and yet to retain the reader's interest.

In a popular account, Winchester cannot take background knowledge for granted. So he must provide potted explanations of the Agricultural and Industrial Revolutions, the social and economic environment of the time, and geology as a developing science with its plethora of characters, who must all be introduced. Inevitably, Smith barely figures in much of this, and yet he and his achievements have to be intruded as much as possible to keep the plot going.

Smith's public rehabilitation can be traced



Rocky representation: part of Smith's geological map of England and Wales.

to the first volume of Charles Lyell's *Principles of Geology* (1830), the book Charles Darwin took on the *Beagle* voyage as his geological bible and in which Smith's achievement was first praised. The following year, Adam Sedgwick, president of the Geological Society, in his Wollaston Medal encomium for Smith, felt "compelled ... to perform this act of filial duty ... and to place our first honour on the brow of The Father of English Geology ... he that gave the plan, and laid the foundations, and erected a portion of the solid walls, by the unassisted labour of his hands".

The title has stuck ever since. Winchester comments, "an injudicious hyperbole, some churlish few will say". In one sense I join the ranks of the churls, in that I would like to know more about how and why the perception of Smith changed so radically within the predominantly middle-class membership of the Geological Society. On this, Winchester has little to say. Why exactly did grantees of the society such as Lyell promote Smith when they did? Was it perhaps to do with competing claims for innovation in geological mapping from abroad? These date back to Johann Lehmann's 1756 section of strata in the Harz mining region of Germany, Georg Fuchsel's 1761 map of Thuringia and the 1811 map of Paris and its environs by Georges Cuvier and Alexandre Brongniart in France, which was directly inspired by Smith's work and by Sir Richard Griffith's work in Ireland. None of these names figure in Winchester's account.

What did Lyell, Sedgwick and the other bigwigs of the society really think of Smith?

I am sure there are clues in their private correspondence. In acknowledging help from historians of science, Winchester admits that "this short book should be thought of simply as the hors d'oeuvre". The chef for the main dish he is referring to is Hugh Torrens, a historian of geology who, according to Winchester, is also writing a biography of Smith. I hope that promise is fulfilled — Smith deserves it after all this time. ■

Douglas Palmer is at 31 Mawson Road, Cambridge CB1 2DZ, UK. He is one of the authors of *The Science Book*, to be published by Cassell this September.

How the old became new

Revolutionizing the Sciences: European Knowledge and its Ambitions, 1500–1700

by Peter Dear
 Palgrave: 2001. 208 pp. £45 (hbk),
 £14.99 (pbk)

Michael Hunter

The concept of a 'Scientific Revolution' — a radical transformation of ideas about the natural world that occurred in the sixteenth and seventeenth centuries — seems to have survived the attacks on it in recent years by revisionists who stress the continuity between old and new ideas in the period. On the other hand, it has become more rather than less difficult to write about the topic. This is partly due to the accumulation of research and partly to the proliferation of different approaches to the subject.

The marxist view of science as being moulded by social forces still exerts a strong influence on ideas about developments in the sixteenth and seventeenth centuries. This period saw the emergence not only of modern science but also of modern capitalism, raising questions about how the two are related.

There was a powerful intellectualist reaction against this view in the postwar years, associated particularly with the historian of science Alexandre Koyré, which stressed the internal dynamics underlying the evolution of scientific ideas. This tradition, too, remains very much alive. More recently, we have seen the rise of cultural history, which looks for subtler social and institutional links between ideas and their context. Peter Dear's brave attempt at a general survey of the way scientific ideas developed in the sixteenth and seventeenth centuries reflects the opportunities and also some of the tensions resulting from these contrasting approaches.

The book is perhaps at its best on the sixteenth century, where Dear clearly illustrates how researchers into all facets of the natural world sought to restore a correct under-

standing of the ideas of antiquity, rather than to go beyond them. Vesalius, Copernicus and others are all convincingly placed in this context. Dear also brings out well the impact of printing on this, as on other, fields of knowledge, and the significance of ideas about magic in stimulating intellectual change.

When dealing with the seventeenth century, the book becomes more technical and in some respects slightly more traditional. Much space is devoted to an exposition of the ideas of a succession of thinkers — from Bacon, Galileo and Kepler to Descartes, Newton and Huygens. Indeed, Dear is perhaps at his strongest in his lucid descriptions of the technical aspects of natural philosophy and the way in which ideas inherited from Aristotle were transformed during the period. He also throws interesting light on the changing criteria used to evaluate natural knowledge, especially the increasing emphasis on experiment. And he does not neglect the biological sciences, even though his main emphasis is on mathematical physics. As a full and accurate account of such matters, this book is the best available, and I would recommend it to anyone.

The later part of the book also describes the context in which ideas developed. It deals in particular with the significance of courtly patronage in supporting intellectuals, and the role of the universities and the new scientific institutions. In this connection, Dear also brings in the rise of museums and botanical gardens, which housed new specimens from exotic locations.

Yet it could be argued that the focus of this part of the book is narrower than in earlier chapters, and readers may be disappointed to find that certain broad issues concerning the changing role of science are dealt with slightly perfunctorily. For example, Dear makes much of recent research that has placed Galileo's ideas in the context of the controversy that existed then between mathematicians and natural philosophers, with mathematicians making increasingly strong claims that they could describe reality rather than purely hypothetical constructs. But Galileo's famous trial is mentioned only briefly, as are other skirmishes between natural philosophers and the Church, such as the burning of Giordano Bruno. Moreover, Dear says virtually nothing about the impetus given to scientific study by rival religious positions, and particularly the vexed issue of the role of 'Puritanism'.

Even the book's dominant — and highly convincing — claim that scientists' ambitions shifted during this time from a desire to understand nature to a wish to control it could have been placed more fully in context. For all Dear's stress on how widely natural philosophers such as Newton or Huygens differed from their ancient and mediaeval precursors in stressing the usefulness of their work, this idea is not fully anchored in its milieu. Without needing to revert to a crude

marxist interpretation, Dear could have explored the broader cultural implications of this characteristic of early modern science at greater length. On the other hand, the extensive bibliography will enable readers to pursue themes that are dealt with tangentially in the book as well as those central to it. Overall, Dear's synthesis is an effective and stimulating one.

Michael Hunter is at the School of History, Classics and Archaeology, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK.

Physics from the inside

The Physics of a Lifetime: Reflections on the Problems and Personalities of 20th Century Physics

by V. L. Ginzburg

Springer: 2001. 513 pp. \$74.95, £51.50

A. M. Bradshaw

A list of the important remaining problems in physics compiled at the dawn of the twentieth century might not have been very long. Several noted physicists, including Albert Michelson, Lord Kelvin and Philipp von Jolly, had already agreed that the cathedral of physics was virtually complete, with only a few turrets and pinnacles to be added, a few roof bosses to be carved. Reality turned out quite differently: quantum mechanics, relativity, elementary-particle theory and cosmology became some of the greatest scientific (and cultural) achievements of the twentieth century.

The same danger lurks in part I of this book by the renowned Soviet physicist V. L. Ginzburg. It is a list, or rather a compendium, of the exciting areas in physics where unsolved problems remain or where interesting developments are likely to occur in the next few years. The first edition of part I (the "Problems" of the subtitle) was published in 1971 and has been updated in successive editions. The present text is a translation of the 1995 edition (with a few addenda and notes), which Ginzburg claims is the last.

The author is aware of the intrinsic problems: "many outstanding discoveries and accomplishments in science come entirely unforeseen and unexpected." Indeed, his 24 topics under the headings 'macrophysics', 'microphysics' (elementary particles) and 'astrophysics' are not a definitive list of the most important problems in physics (although they might resemble this shibboleth, were it to exist). Rather, Ginzburg claims to provide thoughts of "an arbitrary and subjective character" which aim to promote a greater awareness and appreciation of physics, particularly among young people.

The short descriptions of each area are written in lucid, matter-of-fact prose, which is also a compliment to the (unnamed) translators. The level is suitable for an undergraduate, or a layperson with more than a passing acquaintance with physics. But the work is hardly the "popular science publication" Ginzburg claims.

What can be said about his choices and the book's content? The opinions of a great physicist on physics are invariably interesting, and Ginzburg has proved both persistent and perspicacious over the years. A glance at the 1979 edition, which was also translated into English, shows that the importance he attached then to achieving controlled nuclear fusion for energy purposes has not diminished. The search for high-temperature superconductivity has always been an important topic for him, although the celebrated discovery of the effect in a series of rather complex copper oxides did not occur until 1987. Atomic and molecular physics and quantum optics, on the other hand, have never been among Ginzburg's priorities. The recent experiments on Bose-Einstein condensation of gases — almost universally regarded as a very 'hot' topic in physics — receive some attention in a comment. But they are praised mainly as an achievement in experimental physics that was made possible by laser cooling and traps, rather than as a discovery of something unexpected and essentially new.

Current experiments on the interpretation of quantum mechanics are given shorter shrift: "[they] testify to the perfect validity and, one can say, the triumph of quantum mechanics ... critics of quantum mechanics are dissatisfied with the probabilistic nature of part of its predictions. They would, apparently, like to come ultimately to know where each electron goes in a diffraction experiment. There is no hope for this now."

So who is Vitaly Lazarevich Ginzburg, and how has he produced what must be described as a phenomenal overview of physics? Born in 1916, Ginzburg is a theoretical physicist who worked with the Nobel prizewinners I. E. Tamm and L. D. Landau at the Lebedev



Colleagues: (from the right) Ginzburg, Rudolf Peierls, Freeman Dyson and I. E. Tamm in 1956.

Physics Institute in Moscow and later headed its theoretical-physics department. His wide range of interests over some 60 years of research have included nuclear fusion, cosmic rays, radio astronomy, plasma physics, electrodynamics, superconductivity, ferroelectricity and other topics in solid-state research.

He is probably best known for the Ginzburg–Landau theory of superconductivity, which has found applications in other areas of physics. And he worked under Tamm on thermonuclear weapons research. However, because Ginzburg's wife had been jailed for alleged counter-revolutionary activity in 1944 and was subsequently exiled to Gorky (where they met), he was regarded as a security risk and did not accompany Tamm and Andrei Sakharov to the KB-11 laboratory in Sarov, Russia's nuclear weapons design centre, in 1950.

Despite his awareness of Landau's problems in 1938 — he was imprisoned for writing a genuine counter-revolutionary pamphlet — and the fate of other less fortunate colleagues, disillusionment set in only later and is honestly expressed: "I am fully aware of my past grievous delusions (for instance, I realised that Stalin was a foul criminal and murderer only after the Communist Party had revealed that in 1956)."

The middle part of the book contains scientific autobiographical notes and essays on the philosophy of science. And Ginzburg gives his views on questions such as age and productivity, priority in the attribution of discoveries — a topic he returns to several times — and Nobel prizes. He points out that Russian physicists were at a distinct disadvantage in the 1920s and 1930s in being less frequently nominated by their colleagues for the Nobel prize than were their counterparts in the West. For instance, G. S. Landsberg and L. I. Mandelstam discovered the Raman effect just days before the great Indian scientist after which it is named. And although their results were published in the same year, the 1930 Nobel prize went to C. V. Raman alone. (The German physicist A. Smekal, who predicted the effect, was similarly ignored.)

Ginzburg assiduously refers in the book to Vavilov–Cerenkov radiation to emphasize that it had two discoverers. Vavilov, a director of the Lebedev Institute, died before the Nobel prize was awarded to Cerenkov, Tamm and Frank, the latter two having described the effect theoretically. Do we need Nobel prizes in science? On balance, Ginzburg thinks we still do.

The most interesting part of the book, for me at least, is the third part. This consists of short biographies and reminiscences of great Russian contemporaries, or near contemporaries, such as Landau, Tamm, Landsberg, M. V. Keldysh and, above all, Sakharov. The essay on "The Sakharov phenomenon" post-dates most of the other biographical contributions and concentrates on the years of

Science in culture

A platform of humankind

The Venice Biennale, the 49th International Exhibition of Contemporary Art.

Megan Williams

As science becomes ever more present in popular culture and the media, so the way it is depicted in the arts is also changing. In previous decades, science has often been portrayed in art as an ideological weapon turned against itself. But it is now increasingly viewed by artists as a politically neutral realm for exploration.

This year's Venice Biennale, the 49th international exhibition of some of the most electrifying modern art in the world, is proof in point. Science seems to permeate the exhibition's pavilions — perceptions of the human body and mind, questions about social behaviour, our environment, the rhythms of everyday life, and new technologies are probed in a significant proportion of the works on display.

The scientific and psychological exploration ranges from British sculptor Ron Mueck's hauntingly lifelike silicon sculptures of the human body to Hungarian artist Tamás Komoróczy's depiction of obsessive-compulsive disorder. In Komoróczy's installation, the walls of the Hungarian pavilion are covered with a mural composition reflecting the psychological disorder. It consists of computer-generated, rhythmically alternating stripes based on the principle of the 'sampling mix' used in synthesized music, but in this case it is reproducing and manipulating samples of images. Komoróczy also uses video animation and sound corridors, which involve the repetition of and alterations to images and sounds. The result is an artistic experience that replicates obsessive-compulsive disorder in content and form, with its characteristic coinciding and conflicting patterning.

Processes of the mind are also explored in British artist Keith Tyson's installation, *Drawing and Thinking*, which expresses a sense of awe at the mystery of thought. A principal work in this installation, entitled *The Thinker (After Rodin)*, is a mural depiction of a hexagonal tower housing an artificial-life programme that drives its own artificial universe and yet is unable to communicate with the outside world. Tyson's



painting is a commentary on the human thought process — what he calls "the complex process of cognition, or consciousness or self-awareness" originating from, and held within, the human brain. It reflects his amazement at the ineffable quality of human thought — the fact that science is on the threshold of inventing silicon-based machines able to replicate the human cognitive process and yet we are still unable to know exactly what someone else is thinking, nor communicate fully what we ourselves are thinking.

This year's Biennale, whose theme is the 'Platform of Humankind', clearly reflects how the content of contemporary art as well as its mode of expression are now heavily influenced by scientific inquiry. Sound patterning, video imaging, even the use of the olfactory sense as a means of expression, have all become part of the artistic tool-box. As is so evident here, the traditional boundaries between art and science have never been less apparent — to the enrichment of both.

Megan Williams is a freelance journalist based in Rome.

The Biennale runs until 4 November 2001.

opposition and Sakharov's exile to Gorky in 1980. Sakharov's complex personality — characterized by what Ginzburg describes as "apartness" — led to conflicts with his colleagues, although they remained totally loyal to him during his years of petty and humiliating persecution.

What some readers may miss in this book is an account of Sakharov's and, indeed, Ginzburg's contribution to the Soviet thermonuclear-weapons programme. Tantalizingly, Ginzburg writes (actually in 1991) that this information is still classified, but that there had been two key ideas, one from

Sakharov and the other from himself. In a footnote we learn obliquely that Ginzburg's was the use of lithium-6 deuteride as the fusion fuel in the bomb; we now know that Sakharov's contribution was the 'layer cake' configuration. But Ginzburg was not really an insider and he has decided — perhaps wisely — not to update these chapters. What we do learn is that — just as in the United States — some of the finest minds in physics worked on the nuclear-weapons programme.

A. M. Bradshaw is at the Max Planck Institute for Plasma Physics, Boltzmannstrasse 2, 85748 Garching, Germany.

Owen's Parthian shot

Charles Darwin may have had the science, but Richard Owen could write a lethal letter.

Kevin Padian

At first glance, the letter reproduced overleaf appears to be an extraordinary: Sir Richard Owen, the greatest comparative anatomist of the Victorian age, stalwart opponent of transmutation, said to be the only man that Charles Darwin ever hated, is hailing Darwin as "our British 'Copernicus of Biology'" and is apparently recommending him for a statue in the Natural History Museum. But a close reading reveals a very different message. Instead, Owen is using his considerable rhetorical skills to belittle Darwin as much as he is able (under the circumstances of Darwin's recent death), and to discourage any such honour for him.

The cascade of tribute and sentiment that followed Darwin's death in 1882 was effusive and even triumphal. Owen and his transcendentalist supporters had been vanquished on the issue of evolution, and were leaving no intellectual heirs. As Owen notes, Darwin had not convinced everyone of the importance of the mechanism of natural selection to evolution, but he had made it philosophically respectable to accept that the emergence of new species — transmutation — occurs by natural means (Owen's secondary law). Serious investigators would no longer be able to presume divine creation (Owen's primary law) or the direct intervention of an 'intelligent designer'. Darwin's ideas changed the world, and he had been buried with great ceremony in Westminster Abbey, an irony not lost on either his friends or his enemies.

Owen's letter is addressed to Spencer Walpole, a trustee of the Natural History Museum who had served as home secretary in several UK governments. Owen had spent much of 20 years lobbying parliament and the royal family for support to build this museum as a separate, free-standing monument to the science and history of the Earth and its life. But beyond this, Owen brilliantly designed the museum as a monument to himself, his transcendental world view, and the ascendant power of natural history in the influential circles of the time. Now, after succeeding, becoming its first director, and retiring amid fading honours and encomia, Owen is being asked to support the reification of his most bitter rival in its very halls.

In constructing his response to Walpole, Owen knows that he must be cautious. He cannot demean Darwin openly, so he begins with fulsome praise for Darwin's substantial contributions to natural history. But even here omissions and distortions creep in. Darwin had

been asked to accompany Robert Fitzroy on the *Beagle*; he had not solicited the post. A gentleman on board a survey vessel was unusual; Darwin was recommended for his acumen. Robert McCormick was the ship's naturalist but Darwin's greater abilities and the captain's favouritism led to McCormick's early departure, officially on the grounds of ill health.

Owen praises Darwin's field skills (of which Owen had none), his hypothesis of coral-atoll formation — still viable today — and his monographs on "structures and vital actions of plants and animals". But he does not mention Darwin's other great work, *The Descent of Man, and Selection in Relation to Sex*, in which the theory of sexual selection was first proposed, as well as a model for the evolution of humans from less human ancestors. To Owen, both were distasteful and speculative.

Owen's argument then develops in the most delicate way. Jean-Baptiste Lamarck, Owen says, was the first to propose that species could evolve by natural means, and he stresses Lamarck's ideas of the "exercise or disuse of parts of the body": the inheritance of acquired characteristics, to which Darwin and most others largely subscribed, and to which Darwin addressed a chapter in *The Origin of Species*. Owen does not mention Lamarck's vitalistic view, to which hardly anyone subscribed: that organic beings could change by exercising their will on a metaphysical force. Owen allows that most biologists accept that evolution (the "secondary law") has occurred, and that Darwin has forcefully advocated his own mechanism of natural selection to explain it. But this is not the only mechanism of evolution, Owen says. It is not universally accepted. Indeed, he states, Lamarck's ideas are often more applicable — although Owen does not cite cases.

This is why Owen compares Darwin to Copernicus. The Renaissance astronomer's discoveries were valuable to the nascent science of his time, but were later overshadowed by the works of Galileo, Kepler and Newton. So, Owen implies, Darwin's views are similarly rudimentary. The difference between Darwin and Copernicus — Owen cannot resist the further insult — is that Copernicus didn't claim to have the final answer in his field. Darwin, in the end, was not terribly original, and insofar as he departed incautiously from the ideas of others about the 'secondary law by evolution', his work is probably fated to be largely superseded.

But then, Owen seems to strike a positive note. Surely we will learn more in the future about the mechanisms of evolution, and Dar-



Barbed: Owen's letter belittles Darwin (see over).

win's early labours will be recognized by his memorial in Westminster Abbey. As to a statue in the museum ... well, that's done from time to time for individuals who have made extraordinary contributions. And Darwin, as Owen noted, did collect plants and animals and fossils, which he turned over to the "competent officers" to study. (Owen had been one of these "officers", and he mistook the South American fossil mammals so badly that it set back their taxonomy, and the basis for Darwin's ideas about biogeography, for decades.)

And here, the final letdown. Owen predicts that Darwin's work, like Copernicus's, will be eclipsed by later workers. Thus, Owen sniffs, whether scientists of international reputation would find Darwin's contributions worthy of a statue in the museum "may be a question for 'Administration'."

So Owen, like Georges Cuvier in his 'elegy' of Lamarck many decades earlier, managed a Parthian shot at his rival in the guise of a tribute — a masterful and delicate duplicity. How Walpole and his trustees read Owen's veiled attack may never be known. But in the end, it was in vain. Darwin got his statue in the museum, and Owen eventually got his too — although on a different staircase.

Kevin Padian is in the Department of Integrative Biology and Museum of Paleontology, University of California, Berkeley, California, USA. The author wishes to thank J. David Archibald and Angela Milner.

FURTHER READING

- Rupke, N. *Richard Owen: Victorian Naturalist* (Yale University Press, New Haven, CT, 1995).
- Hull, D. L. *Darwin and his Critics* (University of Chicago Press, Chicago, 1973).
- Desmond, A. & Moore, J. *Darwin* (Warner Books, New York, 1991).
- Burkhardt, R. W. *The Spirit of System: Lamarck and Evolutionary Biology* (Harvard University Press, Cambridge, MA, 1977).

Sheen Lodge, Richmond Park, E. Sheen, S.W.
5th November, 1882

Dear Mr. Walpole,

In compliance with your request I have the pleasure to send the following on the subject we last discussed. Charles Darwin has peculiar claims to fitting posthumous recognition of his services to natural science. Of independent means, he devoted himself at the successful termination of his University career to the advancement of natural history. His desire to accompany as naturalist the circumnavigatory expedition of H. M. S. *Beagle* under Captain Fitzroy was granted. The results to his favorite science were equal to, if they did not surpass, those of the naturalists Banks and Solander in the circumnavigatory voyage of Captain Cook. Darwin brought home rich collections in zoology, botany and palaeontology, and liberally made them over to national museums, on the condition of their being described by the competent officers. The results are the richly illustrated quartos, published by the Government, forming with Darwin's own *Notes on the Voyage*, in the well-known 8vo work, the most instructive and exemplary record of the natural-history gains of the circumnavigation. Perhaps the most important and novel researches made during the voyage are those in the nature and growth of coral-formations classified by him as 'atolls', 'barrier-reefs' and 'fringing-reefs', the description and explanation of which Darwin gives in his classical work on *The Structure and Distribution of Coral Reefs* (8vo, 1842).

Since that date he has enriched his favorite science from time to time by monographs throwing most acceptable light on structures and vital actions of plants and animals; they are classical and perennial acquisitions to biology. The guiding principle underlying these works is that advocated in the *Philosophie Zoologique* of Lamarck, on the origin, viz., of species by secondary law, or evolution. But Lamarck's notion of the way of operation of that 'law', viz., by conditions affecting the exercise or disuse of parts of the body, is but partially accepted by Darwin; he substitutes another, a wider, and, as he deems, a truer way of the operation of such 'secondary law', which he sums up under the term 'Natural Selection'.

The great value of Darwin's series of works, summarizing all the evidences of embryology, physiology, palaeontology then accessible, with experiments on the variation of species, is exemplified in the general acceptance by biologists of the 'secondary law by evolution' of the 'origin of species.' As a rule, summaries and monographs now published in natural history are penned under the influence or in acceptance of that 'law'. In this respect Charles Darwin stands to biology in the relation in which Copernicus stood to astronomy. The rejection of the origin of species by primary law, or direct creation, is equivalent to the rejection of the fixity, centrality and supreme magnitude of the Earth; it parallels the substitution of the heliocentric for the geocentric hypothesis. The accelerated progress of natural history under the guidance of 'evolution' resembles that of astronomy under the guidance of 'heliocentricity.'

But the adoption of Darwin's hypothesis of the evolutionary way of work is not general: Lamarck's hypothesis is found in some cases to be more applicable. And so it seems that Darwin parallels Copernicus: save that the latter not only knew not, nor feigned to know, how the planets revolved round the sun.

For that knowledge were requisite the subsequent labours of a Galileo, a Kepler, a Newton. Analogy raises the cheerful hope, if not confident expectation, that the science of living things will also be helped by its Galileo, its Kepler, finally its Newton; and that the way of operation of the 'secondary law originating species' will be as firmly established as is the 'law of gravitation'. Meanwhile our British 'Copernicus of Biology' merits the manifestation of gratitude and the honour which the Empire confers by a Statue in Westminster Abbey.

In the British Museum sculptural memorials have been accorded to meritorious officers; – to Panizzi in relation to the Department of Printed Books; to John Edward Gray, in relation to the Department of Zoology.

Whether the estimation of scientists at home or abroad of Charles Darwin's claims to posthumous honour be met, or their expectations fulfilled, by placing a statue in the Museum of Natural History may be a question for 'Administration'.

Believe me,

Faithfully yours,

Richard Owen

Rt. Hon. Spencer Horatio Walpole, M.P.

Phenomenal fluids

Martyn Poliakoff and Peter King

Heat a liquid in a sealed tube; it seems a simple enough experiment. But it has intrigued chemists and physicists for over 175 years, probably because the results are unexpected, almost counterintuitive.

What you see depends on how much liquid is in the tube. If the tube is almost empty, the liquid evaporates quickly and you are left with a gas at moderate pressure. If it's almost full, the liquid expands rapidly to fill the whole tube and the pressure rises alarmingly. Put in just the right amount, and the meniscus grows faint and disappears abruptly: the contents of this tube have passed through the 'critical point' and become 'supercritical'.

If you heat the tube more slowly, the liquid begins to look opalescent as it approaches its critical point. As the opalescence increases it grows redder and darker and, if you have judged things really well, goes completely black. At the critical point, it becomes perfectly reflecting and you see your own eye staring back at you. Heat it a bit more and the fluid passes back through red to become completely transparent again. Now, perhaps you see why supercritical fluids (SCFs) have fascinated people for so long. A pure substance in a sealed tube turns black merely because of a tiny increase in temperature — in some cases an increase as small as one-hundredth of a degree.

Opalescence arises because the compressibility of the fluid at the critical point is infinitely high. When this happens the microscopic thermal fluctuations that occur

naturally in any fluid become strongly correlated, leading to large-scale, coherent density fluctuations. These large density fluctuations are very effective at scattering light. The high compressibility and the correlated fluctuations also cause the speed of sound to drop to a minimum as the fluid passes through the critical point and the fluctuations attenuate the sound. In 1822, these acoustic effects were exploited unwittingly by Cagniard de la Tour when he heated alcohol in a sealed gun barrel and listened to the musket ball rolling about. Acoustic measurements are still important for studying SCFs: we have been using them to locate the critical points of chemical reaction mixtures.

The whole tube is now filled with a supercritical fluid of uniform density. Technically it's a gas, but it retains some properties of a liquid, including the ability to dissolve many solids. The idea that solids can dissolve in gas is so surprising to some scientists that, on first hearing of it, they fondly imagine that SCFs might dissolve substances that generations of chemists had failed to get into solution. In reality, supercritical CO_2 (scCO_2), the most common SCF, has a solvent power similar to that of petrol — although it does have the unusual ability to dissolve fluorinated organic compounds.

In recent years, industries have used SCFs, particularly scCO_2 , as solvents for applications ranging from the extraction of caffeine, scents and essences to the degreasing of machine components. The handy thing is that the solubility of many compounds in CO_2 changes markedly near the critical point. Lower the temperature or pressure by a small amount and it may well start snowing caffeine or even grease, and the CO_2 can be re-used or released back into the atmosphere.

There is increasing interest in SCFs, particularly scCO_2 , as environmentally acceptable solvents for chemical reactions. But SCFs are attractive to chemists for reasons other than environmental friendliness. SCFs can provide reaction conditions that may lead to enhanced rates or alternative reaction pathways to those offered by conventional solvents. Supercritical water is so reactive that materials ranging from paintshop waste to nerve-gas weapons are rapidly converted into a few harmless oxides, water

Supercriticality

A strange and intriguing state in which solids can dissolve in gases, and liquids can alternate between reflectivity and transparency.

and CO_2 . In many cases, reaction products may be extracted by making tiny reductions in pressure or temperature.

Chemical reactions involve mixtures, and the critical behaviour of mixtures may be far more complicated than that of a pure substance,

with two or more liquid phases as well as the gas phase. The challenge in controlling the reaction often lies in characterizing and using this complex behaviour.

Strictly speaking, the critical point of a mixture is reached when the liquid and gas phases coalesce and the densities of the two phases are equal at the moment of coalescence. In practice, any homogeneous mixture at a temperature above the critical point of the pure solvent is termed 'supercritical'.

Returning to the sealed tube, what happens when our simple SCF is rapidly cooled? While it is supercritical, the fluid has a uniform density in the tube; therefore, as it cools through the critical point and separates into two phases, gas bubbles and liquid droplets form with equal probability throughout the tube. The droplets begin to fall and the bubbles rise. A veritable storm erupts — quite different from the gradual disappearance of the meniscus when the liquid is heated. In short, a second beautiful phenomenon for the price of one!

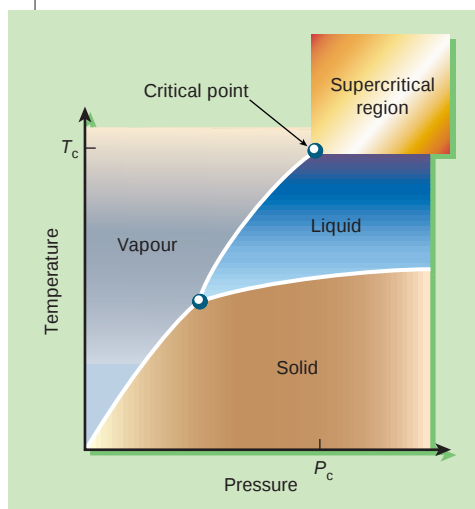
Martyn Poliakoff is in the School of Chemistry and Peter King is in the School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK.

FURTHER READING

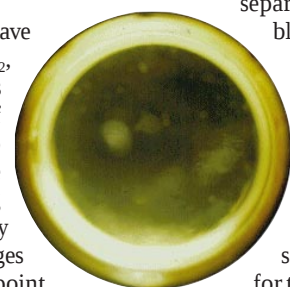
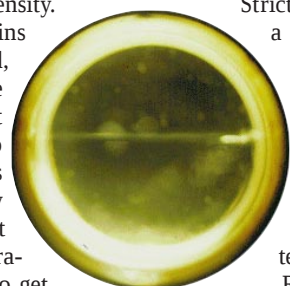
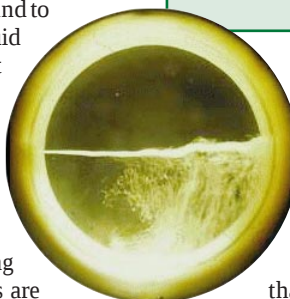
McHugh, M. A. & Krukonis, V. J. *Supercritical Fluid Extraction* (Butterworth-Heinemann, Boston, 1994).
Jessop, P. G. & Leitner, W. (eds) *Chemical Synthesis Using Supercritical Fluids* (Wiley-VCH, Weinheim, 1999).
Noyori, R. (ed.) *Chemical Reviews* **99**, no. 2 (1999).
Adam, D. *Nature* **407**, 938 (2000).

WEBLINK

<http://www.nottingham.ac.uk/supercritical>



Critical conditions: above, phase diagram for a pure substance showing the critical pressure, P_c , and temperature, T_c . Above right, the effect of heating CO_2 rapidly through the critical point from liquid (top) to supercriticality (bottom).



Lost City found

Karen L. Von Damm

Chemical and heat exchange at vents on deep ocean floors has a large influence on marine chemistry. The discovery of a spectacular new type of venting system has given the story another twist.

One of the revelations of twentieth-century science was that the deep oceans contain vast mountain chains, rising from the sea floor. Here, as a consequence of plate tectonics, hot material rises from within the Earth's mantle, generating new ocean crust and forming mid-ocean ridges. Over the past 25 years, it has emerged that because of hydrothermal activity there are places on these ridges where large-scale chemical and fluid exchange occurs between the solid Earth and the ocean, and unusual animal communities thrive. The amazing undersea images of these hydrothermal fields, especially of sulphide-rich 'black smoker chimneys' and lush chemosynthetic animal communities, came as a complete surprise.

Now we have another reminder of how little we know about the part of our planet's surface — over 60% of it — that is occupied by sea floor. As Kelley *et al.* describe on page 145 of this issue¹, last December they discovered a new type of seafloor hydrothermal field, using the research vessel *Atlantis* and the submersible *Alvin*. The field, called Lost City, is in the North Atlantic, off the Mid-Atlantic Ridge (see map on page 146). It is visually spectacular, and has a style of venting and a geographical setting that distinguish it from any previously discovered system of hydrothermal vents.

Lost City is characterized by massive white structures, up to 60 m high, rather than black smokers (Fig. 1). It is situated near a 'slow-spreading' mid-ocean ridge (the significance of which is discussed below), but not on the ridge itself. Instead, it is 'off-axis', nearly 15 km away from the ridge axis. The distinction between off- and on-axis is drawn at a crustal age of about one million years; because mid-ocean ridges are spreading, the further the crust from the axis, the older it is. In the case of Lost City, 15 km corresponds to 1.5 million years.

Since the discovery of undersea hydrothermal sites, the overriding goal has been to understand the influence of hydrothermal activity on ocean chemistry. This matters because ocean chemistry largely determines how well the oceans can buffer changes in the climate system, and also what kind of life can thrive in it — as has been the case throughout Earth's history. The oceans may well have been the place where life on Earth originated.

Lost City is a different type of vent system to those previously known. Further research

will be aimed at tackling three questions in particular. First, what is the relative importance of off-axis, compared with on-axis, activity in terms of overall hydrothermal

activity on the sea floor? Second, what are the differences between fluids generated by reaction with rocks that are typically found much deeper in the oceanic crust (peridotites or

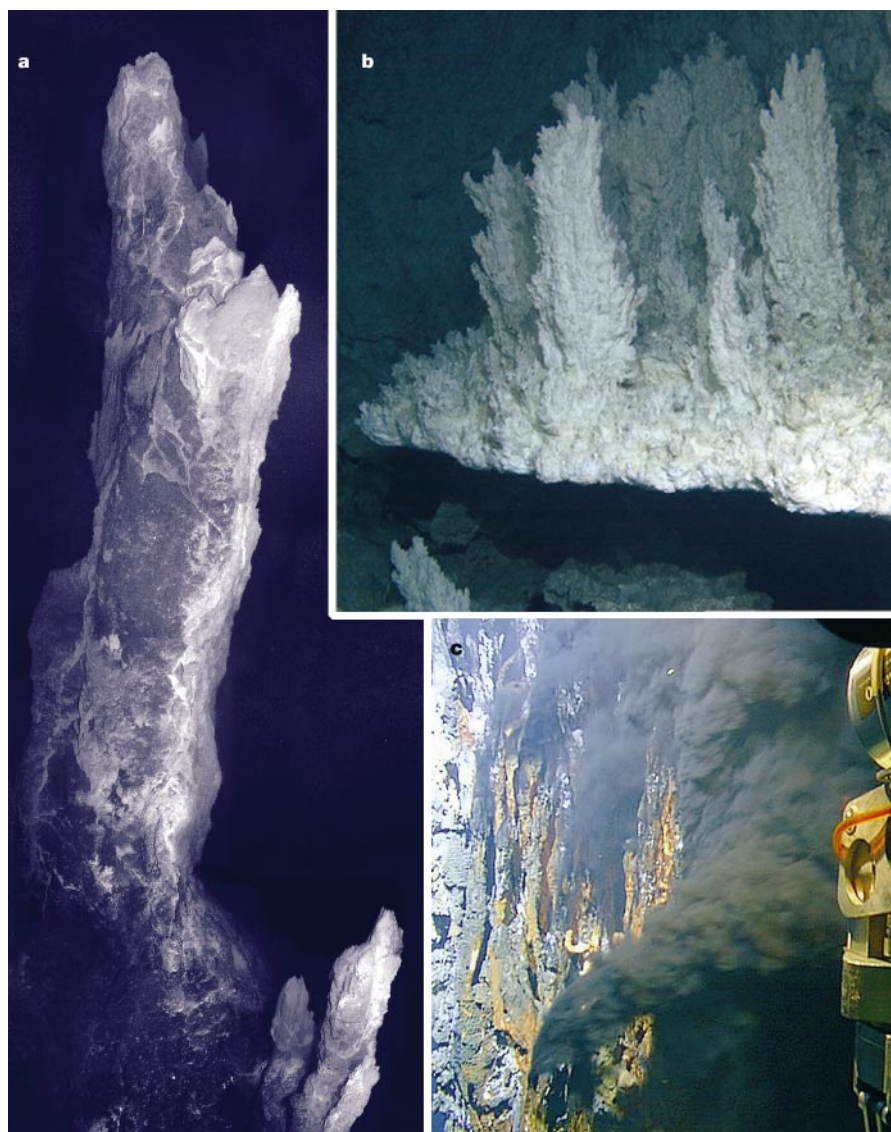


Figure 1 Hydrothermal vents in white and black. a, Mosaic of a structure, composed of carbonate and over 10 m in height, typical of those at Lost City identified by Kelley *et al.*¹. Water depth here is 700–800 m. b, A carbonate flange (ledge) at the same location. Some of these flanges are host to microbial communities that thrive in the warm (40–75 °C) vent waters. Lost City lies 15 km off the axis of the slow-spreading Mid-Atlantic Ridge. Because of the high pH and low temperature of the hydrothermal fluids, their iron concentrations should be low. So mineral structures develop from carbonates and hydroxides, which tend to be light coloured. c, Brandon Vent, a black smoker on the southern East Pacific Rise at a depth of 2,834 m. Apart from being especially hot (405 °C), Brandon is typical of fast-spreading mid-ocean ridges. The acid and iron- and sulphide-rich hydrothermal fluids turn black as they mix with the cool (2–5 °C) alkaline sea water surrounding the vent. All images were taken from the submersible *Alvin*.

M. ELLEND

K. L. VON DAMM/M. D. LILLEY

ultramafics), as seems to be the case at Lost City, and those that result from reaction with the basalts that are present closer to the surface? Third, what is the nature of the microbial communities that might exist at Lost City?

It has long been known that the amount of heat lost by convection on the flanks of mid-ocean ridges is much larger than that lost from the axis itself². The ratio is roughly 70:30, and the reason is that, although the axis is hotter, the flanks have a much larger area. But chemical concentrations (for instance, of iron) tend to be much higher on the axis than away from it. So debate over the relative contributions of off-axis and on-axis fluxes has centred on whether the chemical anomalies found on the axis are great enough to outweigh the greater heat (and probably water) flux on the flanks. The trouble is that off-axis venting is difficult to locate because the lower temperature of the fluids means that certain indicators in the water column (temperature, particles or salinity) or on the sea floor (large constructional features, or animal communities) are absent. Lack of knowledge of off-axis vents has also hampered attempts to estimate the overall chemical flux to the ocean through hydrothermal activity and its importance, relative to the contribution of rivers, to ocean chemistry³.

The mid-ocean ridge system is huge, around 60,000 km in total length, and different ridges in the system spread at different rates. The Mid-Atlantic Ridge, with which the Lost City is associated, is slow-spreading (at a full opening rate of only 2.5 cm yr⁻¹). This compares with up to 15 cm yr⁻¹ on fast-spreading ridges such as the East Pacific Rise. The previously identified sites^{4,5} of off-axis hydrothermal activity are all close to ridges spreading at intermediate rates, as opposed to the slow-spreading location of the Lost City. The topography of slow-spreading ridges is especially rugged, which further hampers the search for hydrothermal sites.

The spreading rate of a mid-ocean ridge is relevant to hydrothermal fluids because a determinant of the fluids' nature is the rock with which they react in passing through the ocean crust. Faster-spreading ridges are dominated by basaltic (mafic) rock types; on slower-spreading ridges, however, fluids may also react with deeper-lying peridotites (ultramafics), leading to the formation of serpentinites and thereby to different fluid compositions. Most of the known sea-floor hydrothermal sites are found on basalt. But on slower-spreading ridges, the ultramafic parts of the oceanic crust are often exposed or closer to the sea floor, and venting fluids may be derived by reaction with this rock type.

This is what seems to be happening at Lost City, where the vent fluids are quite cool (40–75 °C) and alkaline (pH 9 and above).

Another example of seafloor hydrothermal fluids derived from serpentinites is at the Mariana trench and island arc system in the western Pacific⁶. Here the fluids are cooler than ambient sea water and likewise have higher pH values. A similar situation occurs on land — hot springs issuing from serpentinite substrate have extraordinarily high pH values⁷.

Two other places, both on the Mid-Atlantic Ridge, have been discovered where the hydrothermal fluids seem to be derived from ultramafic reactions. They are the Logatchev and Rainbow sites, at 14° 45' N and 36° 16' N, respectively^{8,9}. The chemistry of the fluids at these sites is fundamentally different from those at basaltic sites (they are high in iron, calcium and transition metals, but low in silica). In contrast to the Lost City, however, Logatchev and Rainbow have high temperatures (350 °C or above) and very acidic pH values.

What about the microbial communities at seafloor hydrothermal sites? Such communities depend on chemical energy, often in the form of reduced gases such as hydrogen and hydrogen sulphide, for their existence. Anomalously high concentrations of methane, another reduced gas, have been found in the water column near the Logatchev site¹⁰, but no discrete methane source has ever been located. Another unusual aspect is the hydrogen content of the fluids issuing from ultramafic source rocks, which is greater than that of most basalt-derived hydrothermal fluids. It is not as high as the hydrogen content in vent fluids immediately after volcanic perturbation, but such perturbations are only transient events. These high levels of gas led to speculation as to the importance of ultramafic sites as microbial incubators and in the evolution of

life on Earth. At 40–75 °C, the hydrothermal fluids at Lost City present a much more hospitable temperature range for life than the 350 °C or so of the Rainbow and Logatchev sites. Kelley *et al.*¹ have done some culturing work with microbes from the site. But it will take detailed DNA analyses to see whether they are unique, and if they are more like the microorganisms thought to have existed on the early Earth than are known microbial communities.

Even in the twenty-first century, unexpected discoveries in our oceans are capturing our imagination and forcing us to revise our ideas about processes on Earth. We cannot yet be sure what the chemical controls on the hydrothermal fluids at Lost City may be, or how they might be affected by biological activity. However, as much of the global ridge-crest system can be characterized as slow-spreading, with ultramafic source rocks, we can be sure that Lost City will provide new ways of thinking about seafloor hydrothermal activity in general and its effect on ocean chemistry. ■

Karen L. Von Damm is at the Institute for the Study of Earth, Oceans and Space, University of New Hampshire, 366 Morse Hall, 39 College Road, Durham, New Hampshire 03824-3525, USA.
e-mail: kvd@cisunix.unh.edu

1. Kelley, D. S. *et al.* *Nature* **412**, 145–149 (2001).
2. Stein, C. A., Stein, S. & Pelayo, A. M. in *Seafloor Hydrothermal Systems: Physical, Chemical, Biological and Geological Interactions* (eds Humphris, S. E. *et al.*) 425–445 (Am. Geophys. Union, Washington DC, 1995).
3. Elderfield, H. & Schultz, A. *Annu. Rev. Earth Planet. Sci.* **24**, 191–224 (1996).
4. Maris, C. R. P. & Bender, M. L. *Science* **216**, 623–626 (1982).
5. Mottl, M. J. *et al.* *Geology* **26**, 51–54 (1998).
6. Fryer, P. *et al.* *EOS* **68**, 1534 (1987).
7. Barnes, I. *et al.* *Contrib. Mineral. Petrol.* **35**, 263–276 (1972).
8. Bogdanov, Y. A. *et al.* *BRIDGE News* **9**, 9–13 (1995).
9. Donval, J. P. *et al.* *EOS* **78**, F832 (1997).
10. Charlou, J. L. *et al.* *EOS* **78**, F831 (1997).

Cognitive neuroscience

Bold insights

Marcus E. Raichle

Functional magnetic resonance imaging tracks changes in oxygen levels in the brain in response to different stimuli. The neural basis of these changes has, at last, been pinned down.

Over the past decade, research in the field of cognitive neuroscience has grown exponentially. To probe the mysteries of the human brain, scientists combine the experimental strategies of psychology with techniques that enable them to examine how brain activity supports mental processes. Leading the way are two techniques for imaging brain function: positron-emission tomography (PET), and magnetic resonance imaging, or functional magnetic resonance imaging (fMRI) as it is now called. fMRI is the main tool for imaging normal brain activity.

fMRI effectively measures the level of oxygen in the blood in the brain, which varies because changes in brain activity are invariably accompanied by changes in blood flow, causing the oxygen level in the blood to rise in the brain region concerned. The robust empirical relationship between changes in brain activity and blood flow has fascinated scientists for well over a century¹, but its neural basis has, until now, remained largely unknown. Despite this, researchers have assumed that the signals they obtain from fMRI are related to actual changes in neuronal activity. On page 150 of this issue,

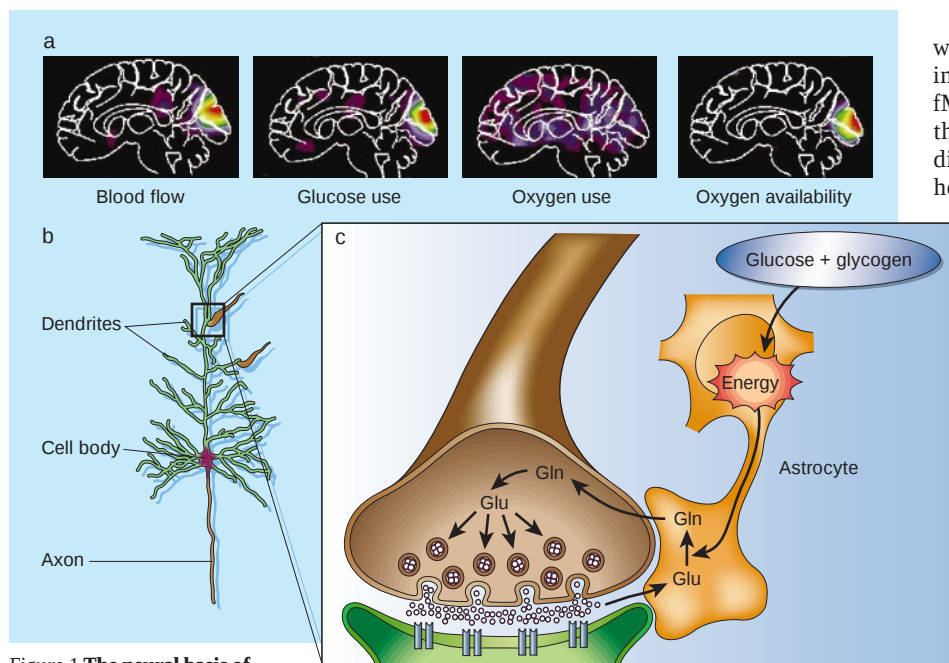


Figure 1 The neural basis of functional magnetic resonance imaging (fMRI). a, Viewing a stimulus such as a checkerboard produces marked changes in the areas of the brain that respond to visual stimuli, as seen in these positron-emission tomographic (PET) images. These changes include increases in glucose use and blood flow that are much greater than those in oxygen consumption. As a result there is an increase in the oxygen level in those areas (supply exceeds demand). PET is usually used to monitor blood flow. fMRI detects the changes in oxygen availability as a local change in the magnetic field. The resulting fMRI signal is a 'blood-oxygen-level-dependent' (BOLD) signal. b, As Logothetis *et al.*² show, these metabolic and circulatory changes are driven by electrical potentials arising from the input to, and information processing within, the dendrites of neurons. c, An attractive explanation for the BOLD signal invokes the preferential use of glycolysis in nearby non-neuronal cells (astrocytes) to handle an increase in the release of the neurotransmitter glutamate (Glu), which must be converted to glutamine (Gln) before it is returned to the neuron. Glycolysis consumes glucose to produce energy, but does not require oxygen.

Logothetis and colleagues² confirm this, in an experimental *tour de force* that represents the first comprehensive look at the relationship between the fMRI signal and the underlying neural activity.

Logothetis *et al.*'s results emerge from their pioneering development of fMRI, whereby both fMRI and various measures of the electrical activity of neurons are carried out simultaneously in monkeys³. The experiments² required monkeys to sit in an MRI scanner while viewing checkerboard patterns. The results show that a spatially restricted increase in the fMRI signal — in this instance in the brain's visual cortex — directly reflects an increase in neural activity.

With their electrical recording techniques, the authors analysed several aspects of neuronal activity, and were able to distinguish between action potentials and local field potentials. Action potentials are the all-or-nothing firing rates of individual neurons and groups of neurons; they occur immediately after a stimulus has been presented, and reflect neuronal output. Local field potentials are the more slowly varying electrical potentials that arise from the input to, and integrative processes within, neurons. The major determinant of the fMRI signal

turned out to be the local field potentials. So, activation of an area of the brain, as seen by fMRI, predominantly reflects the input to that area and the corresponding changes in information processing, rather than output from the area. Notably, much earlier work, using autoradiographic measurements of glucose consumption by different brain areas in rats, anticipated exactly this⁴.

The results hold lessons for both cognitive neuroscientists and cellular neurophysiologists. Cognitive neuroscientists using fMRI must be aware that the signal-to-noise ratios for neural signals recorded directly from the brain are much greater than the accompanying fMRI signal. So the absence of an fMRI signal does not necessarily mean that no information processing is going on in a particular area of the brain. For the neurophysiologists, who seek to understand the biochemical and biophysical processes underlying neural activity, the absence of action potentials must not be interpreted as the absence of information processing. Before neurophysiologists and cognitive neuroscientists can cooperate as needed, both groups must have a working knowledge of the relationships between neuronal activities and fMRI signals.

More fundamentally, Logothetis *et al.*'s work² comes at a time of increasing interest in the general cell biology underlying the fMRI signal. The nature of the signal and the physiological research preceding its discovery have forced us to reconsider how changes in energy production support increases in neuronal activity. At rest the human brain uses up 20% of the oxygen needed by the body, although the brain accounts for less than 2% of the body's mass. The oxygen is used in breaking down glucose to supply the brain with energy.

Surprisingly, however, brief increases in the activity (and energy requirements) of particular brain regions are accompanied by increases in blood flow and glucose consumption that far exceed the increase in oxygen consumption⁵. This is because glucose is being broken down anaerobically, by the rapid process of glycolysis, to supply energy. As a result there is an increase in the amount of oxygen in the blood nearby — supply exceeds demand (Fig. 1a). fMRI is sensitive to changes in the oxygen content in blood, and so detects the 'blood-oxygen-level-dependent' (BOLD) signal⁶.

Recent work has offered a tentative explanation for the local increase in glycolysis^{7,8} (Fig. 1b, c). The communication between neurons occurs at synapses, and requires the release of neurotransmitters from a presynaptic neuron and their detection by a postsynaptic nerve cell. The amino acid glutamate is the main excitatory neurotransmitter in the brain. After it is released, glutamate needs to be removed promptly from the synapse, and this occurs by uptake into an adjacent non-neuronal cell — an astrocyte. There, glutamate is converted to glutamine before being returned to the neuron and recycled. The energy needed for processing glutamate is provided by glycolysis (glucose breakdown without oxygen), using glucose obtained from blood and, during sudden increases in activity, from a unique glycogen storehouse in astrocytes.

In other words, the blood oxygen level rises because of an increase in the processing of glutamate in astrocytes after excitatory neurotransmission. More specifically, Logothetis *et al.*'s results² mean that we can with some confidence ascribe the BOLD signal to a change in local field potentials in postsynaptic neurons that occurs as a result of changes in excitatory neurotransmission. These alterations in local field potentials probably occur largely in postsynaptic extensions (dendrites) that are associated with local information processing. It is not yet known, however, whether the BOLD signal can also be related to a change in use of other neurotransmitters, such as GABA,

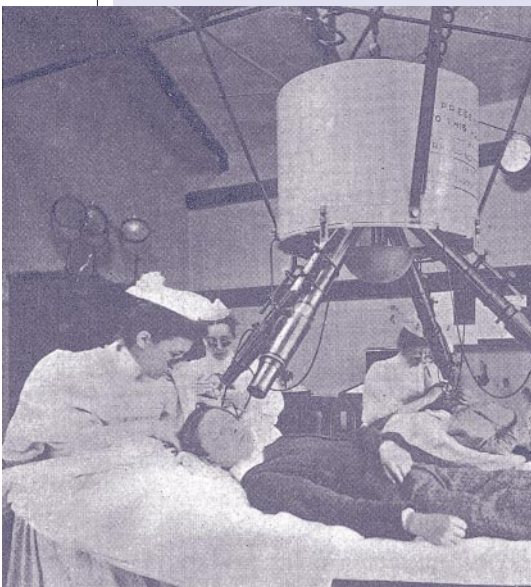


100 YEARS AGO

The treatment of disease by light.

The drawbacks to the treatment are, first, the length of time which a severe case takes, and, secondly, the cost. Not only is there the cost of the electric light and the necessary maintenance, but every patient has to be attended by a nurse. At the London Hospital it has been found that it costs about 400l. or more a year to run one lamp, so that the light department there necessitates an expenditure of 1200l. a year. It is, therefore, gratifying to find that Mr. Alfred Harmsworth has come forward and endowed one lamp by a munificent gift of 10,000l.

From *Nature* 11 July 1901.



50 YEARS AGO

Normally, when an observer is making colour comparisons, he is in the erect position and with normal vision both eyes exhibit similar colour sensitivities. I have observed, however, that if the observer is in the prone position lying on one side, a gradual difference between the colour response of the two eyes develops. After a few minutes the lower eye becomes markedly red-sensitive compared with the upper eye. If now the observer lies on his back, the two eyes gradually return to equality of colour response. By turning on to the opposite side, the eye formerly red-sensitive will be uppermost and will then gradually develop blue-sensitivity compared with the lower eye.

From *Nature* 14 July 1951.

the brain's main inhibitory neurotransmitter.

After a century of research, we still do not know how or why blood flow increases during neuronal activation. It does not seem to reflect an increased need for either oxygen⁹ or glucose¹⁰. But, thanks to the work of Logothetis *et al.*, cognitive neuroscience can move forward with greater confidence in the knowledge that changes in blood flow and oxygen levels do represent definable alterations in neuronal activity.

Marcus E. Raichle is in the Division of Radiation Sciences, Mallinckrodt Institute of Radiology, Washington University Medical Center, 4525 Scott Avenue, St Louis, Missouri 63110, USA.

e-mail: marc@npg.wustl.edu

1. Raichle, M. E. in *Brain Mapping: The Systems* (eds Toga, A. W. & Mazziotta, J. C.) 33–75 (Academic, San Diego, 2000).
2. Logothetis, N. K., Pauls, J., Augath, M., Trinath, T. & Oeltermann, A. *Nature* **412**, 150–157 (2001).
3. Logothetis, N. K., Guggenberger, H., Peled, S. & Pauls, J. *Nature Neurosci.* **2**, 555–562 (1999).
4. Schwartz, W. J. *et al. Science* **205**, 723–725 (1979).
5. Fox, P. T., Raichle, M. E., Mintun, M. A. & Dence, C. *Science* **241**, 462–464 (1988).
6. Ogawa, S., Lee, T. M., Naycik, A. S. & Glynn, P. *Magn. Reson. Med.* **16**, 9–18 (1990).
7. Shulman, R. G., Hyder, F. & Rothman, D. L. *Proc. Natl Acad. Sci. USA* **98**, 6417–6422 (2001).
8. Magistretti, P. J., Pellerin, L., Rothman, D. L. & Shulman, R. G. *Science* **283**, 496–497 (1999).
9. Mintun, M. A. *et al. Proc. Natl Acad. Sci. USA* **98**, 6859–6864 (2001).
10. Powers, W. J., Hirsch, I. B. & Cryer, P. E. *Am. J. Physiol.* **270**, (Heart Circ. Physiol. **39**), H554–H559 (1996).

High-energy physics

Disappearing dimensions

Joseph D. Lykken

Some theories of high-energy physics require extra spatial dimensions, beyond the three we know. A radical proposal turns this idea on its head, and suggests that dimensions may disappear at higher energies.

What is space? Where did the dimensions of our physical world come from? Philosophers since Aristotle have been flummoxed by these ancient questions. Now the particle physicists are having a go — turning philosophy into scientific hypotheses with testable consequences. Two groups of theorists^{1,2}, one based at Harvard University and the other at the Fermi National Accelerator Laboratory in Illinois, have suggested a concrete mechanism for how dimensions of space can come into being, and even disappear.

Aristotle got the first step right: to understand what a spatial dimension is, you need to think about motion. The convincing evidence that we live in three spatial dimensions is that we can move in three independent ways. Objects look three-dimensional because light (composed of particles called photons) moves in three dimensions, obeying three-dimensional laws of optics, and thus of perspective. Any question about dimensionality always boils down to a question about particles and their motion.

In a solid material like a crystal, atoms are held rigidly together, forming a regular lattice of point-like locations in three-dimensional space. Most of the electrons in that material are tied to one particular atom — they effectively live in a world of zero spatial dimensions, because their motion is completely constrained. But in some materials, such as metals, there are residual forces that allow some electrons to hop from one atom to another. Depending on the material, these 'hopping interactions' may allow motion only along a line, or only in a planar surface, or through the full three dimensions of the solid. So, for electrons in

a solid, dimensionality effectively depends on forces.

The provocative models of Arkani-Hamed *et al.*¹ and Hill *et al.*² extend this analogy to elementary particles in a vacuum. Put the Universe under a powerful enough microscope, they say, and you will find that space itself is a lattice, an array of discrete points. Elementary particles, such as electrons, quarks or photons, fundamentally inhabit only a single point. To move, there must be a force — a hopping interaction — that destroys the particle at one point in space and creates a copy of it at a neighbouring point. No force, no motion; no motion, no dimension.

The Harvard and Fermilab theorists have created this microscopic picture of discrete space and hopping particles using simple models in which 'gauge' forces similar to three of the fundamental forces seen in nature — electromagnetic, weak and strong interactions — induce the hopping. Now here comes the tricky part. Gauge forces vary in strength according to the energy involved in the physical process. Electromagnetic forces, for example, get stronger at higher energies, whereas the strong nuclear force between quarks gets weaker at higher energies. In the models created by the two groups of theorists, the hopping interactions actually turn off at high energies, thereby reducing the number of spatial dimensions. Arkani-Hamed *et al.*¹, with postmodernist tongue-in-cheek, call this 'deconstructing dimensions'.

The punchline is that, in the high-energy environment of the early Universe, there may have been no spatial dimensions at all. Dimensionality itself may be a low-energy

phenomenon, which emerged as the Universe cooled down.

This idea is a bit of a stretch even for the *cognoscenti* of string theory, the branch of high-energy physics that attempts a unified description of all fundamental interactions, including quantum gravity. In string theory, space itself — point, lattice or continuum — is supposed to emerge from a more fundamental substrate of wiggling strings. But in string theory there are more spatial dimensions visible at higher energies, not fewer. String theory, for its own consistency, requires ten spatial dimensions. The extra seven dimensions, it is assumed, are hard to see because they are 'curled up' to microscopic size. The number of spatial dimensions thus effectively increases with energy, because studying particle interactions at high energies is equivalent to probing matter at microscopic distance scales. Indeed, inspired by string theory, particle physicists are looking for evidence of extra spatial dimensions in experiments at high-energy particle colliders.

The new models of deconstructed dimensions do not include gravity, and so

cannot yet vie with string theory as a complete description of the world at high energies. More than that, the existing models are really toys, not proper theories of anything. It could well turn out that there is less here than meets the eye. On the other hand, particle theorists have a strong bias that the laws of physics should become simpler at higher energies — and what could be simpler than a world of zero dimensions?

These ideas are extremely speculative, but in a generic way they should be testable. For example, the same experiments that are looking for extra dimensions using particle colliders might turn up evidence that quarks, electrons, gluons or photons are moving in fewer dimensions at the highest energies that can now be produced. The era of post-modern physics may be upon us. ■

Joseph D. Lykken is in the Theoretical Physics Department, Fermi National Accelerator Laboratory, Batavia, Illinois 60510-0500, USA.

e-mail: lykken@fnal.gov

1. Arkani-Hamed, N., Cohen, A. G. & Georgi, H. *Phys. Rev. Lett.* **86**, 4757–4761 (2001).
2. Hill, C. T., Pokorski, S. & Wang, J. *Phys. Rev. D* (in the press). <http://arxiv.org/abs/hep-th/0104035>

Palaeontology

Return to the planet of the apes

Henry Gee

Fossil evidence of human evolutionary history is fragmentary and open to various interpretations. Fossil evidence of chimpanzee evolution is absent altogether.

Discoveries of fossil hominids are like buses: nothing for a while, then three come along at once. Earlier this year, Leakey *et al.*¹ announced *Kenyanthropus platyops*, a 3.5-million-year-old creature with a disconcertingly modern-looking face, given its otherwise primitive cranium. At the same time, another team^{2,3}, also working in Kenya, described remains of a new species, *Orrorin tugenensis*, which at 6 million years old is possibly the earliest known hominid. These discoveries were discussed in accompanying News and Views articles^{4,5}.

In papers beginning on page 175 of this issue, Yohannes Haile-Selassie and colleagues now describe hominid specimens⁶ and palaeoenvironments⁷ from Ethiopia dated at between 5.2 and 5.8 million years old. The hominids are assigned to *Ardipithecus ramidus kadabba*, an archaic subspecies of *A. ramidus*, an early hominid previously discovered⁸ in 4.4-million-year-old sediments in Ethiopia. The designation of *A. r. kadabba* as a subspecies will be controversial. But all concerned agree that both *Orrorin tugenensis* and *A. r. kadabba* are primitive, and they are thought to lie in the family tree close to the point at which the ancestries of

extant chimpanzees and humans diverged. Their phylogenetic position is thus pivotal. Definitive resolutions of the status of these creatures could reveal much about the nature, lifestyle and behaviour of the most recent common ancestor of humans and chimpanzees, and the course of human evo-

lution in general. Whether this potential can be fulfilled is another question entirely.

Their respective discoverers claim that both *A. r. kadabba* and *Orrorin* were bipedal. When *A. ramidus* was first described⁸, bipedality was one of the few features that marked it as a hominid. But *A. r. kadabba* and *Orrorin* are more primitive still, raising the question of whether bipedality is a diagnostic hominid trait. In other words, bipedality, as an habitual form of locomotion, might have occurred in lineages of apes that are now extinct. This idea has found support, albeit controversially, in the claim that *Oreopithecus bambolii*, an ape that lived 7–9 million years ago on an isolated island that is now Tuscany, was bipedal to some extent — and yet this creature is thought to have become bipedal independently and was only distantly related to hominids⁹.

The idea that bipedality was once more widespread than its current humans-only distribution has several implications. First, one would be forced to consider that the ancestors of chimpanzees as well as of hominids were bipedal, and that the distinctive knuckle-walking habit of living chimpanzees is a secondary acquisition. This challenges the controversial idea^{10,11} that the most recent common ancestor of chimpanzees and humans was capable of knuckle-walking.

Second, it is possible that some of these fossils might not be hominids at all (see Box 1 for a guide to the terminology). After all, most researchers agree that the most recent common ancestor of humans and chimpanzees lived around 5–6 million years ago (but see ref. 12), so some of the fossils currently described as hominid might be more akin to chimpanzees, or may represent an entirely extinct offshoot of the ancestry of hominids and chimpanzees — a cousin of the latest common ancestor, if you like.

Moreover, it remains the case that although hominid fossils are famously rare,

Box 1 Hominid and hominin

A 'hominid' is a member of the family Hominidae, which classically includes all creatures, living and extinct, that are more closely related to *Homo sapiens* than to the extant chimpanzees (*Pan troglodytes* and *P. paniscus*), the closest living sister taxon to *Homo*. Senut *et al.*² and Haile-Selassie⁶ use the term 'hominid' in this sense, and so, for consistency, I have done so in the main article here. This classical solution is, however, more problematic for the great

apes — chimpanzees, gorillas (*Gorilla*) and orang-utans (*Pongo*) — which are lumped together in the family Pongidae. The problem is that some of these creatures (chimps and gorillas) are more closely related than others (orang-utans) to humans, in which case Pongidae is not a 'natural' group. One solution is to elevate chimps, gorillas and orang-utans each to their own families. Another is to extend the family Hominidae to include great apes as well as

humans and their immediate, extinct relatives, classifying humans and chimps in a subfamily (Homininae) and demoting hominids (in the old sense) to the subcategory of tribe (the Hominini). This is why Leakey *et al.*¹, using this new terminology, describe as 'hominins' what others continue to refer to as 'hominids'. 'Hominin', therefore, is not necessarily a misprint or a gratuitous attempt to bemuse the unwary.

H.G.

the chimpanzee lineage has no fossil record whatsoever. One explanation has been that chimpanzees have always lived in forested environments, and that forest creatures are rarely preserved as fossils. Hominids only become (relatively) abundant as fossils after they moved from forests to more open habitats. However, this argument is turned on its head by strong evidence that *Orrorin*³ and *Ardipithecus*⁷ lived in woodland. The fossils of animals such as monkeys and small antelopes found alongside the hominids, as well as palaeobotanical and isotopic evidence, suggest that *Ardipithecus* lived in a relatively well-forested and high-altitude environment. Indeed, this creature may have been confined to such a habitat: as Wolde-Gabriel and colleagues⁷ show, searches for early hominids in geological settings indicative of the open-country habitats associated with later hominids were less rather than more likely to produce results. So it may be that hominids were woodland creatures until about 4.4 million years ago^{8,13}.

Given that chimpanzees today live in environments rather like those inhabited by *Ardipithecus* and *Orrorin*, could it be that at least one of these early hominids is actually more akin to chimpanzees? Questions have been raised about the status of *Orrorin* as a hominid^{5,6}. For their part, Senut and colleagues² defend the hominid status of *Orrorin* and propose that *Ardipithecus* is an ancestor of chimpanzees. But they do not discuss the implications of this view for the history of chimpanzees, as distinct from that of hominids.

Sadly, I doubt that the status of these

creatures can be resolved to general satisfaction. Some researchers have suggested¹⁴ that the dental and skeletal traits conventionally used as the basis for hominid systematics are unreliable guides for reconstructing evolutionary history, in that the phylogenies created using these traits differ from those based on molecular information from living primates. Given that bones and teeth are, for practical purposes, all there is to go on, uncertainty is likely to reign for some time, leaving the nature of the latest common ancestor — and the general course of early hominid evolution — as mysterious as ever.

Is the outlook completely gloomy? Perhaps not. The accumulating data on palaeoenvironments should at least improve our understanding of the lives and times of early hominids (and perhaps of early chimps), even though the evolutionary relationships remain murky. ■

Henry Gee is a senior editor at Nature.

e-mail: h.gee@nature.com

1. Leakey, M. G. *et al.* *Nature* **410**, 433–440 (2001).
2. Senut, B. *et al.* *C. R. Acad. Sci. Paris* **332**, 137–144 (2001).
3. Pickford, M. & Senut, B. *C. R. Acad. Sci. Paris* **332**, 145–152 (2001).
4. Lieberman, D. E. *Nature* **410**, 419–420 (2001).
5. Aiello, L. C. & Collard, M. *Nature* **410**, 526–527 (2001).
6. Haile-Selassie, Y. *Nature* **412**, 178–181 (2001).
7. WoldeGabriel, G. *et al.* *Nature* **412**, 175–178 (2001).
8. White, T. D., Suwa, G. & Asfaw, B. *Nature* **371**, 306–312 (1994).
9. Köhler, M. & Moyà-Solà, S. *Proc. Natl Acad. Sci. USA* **94**, 11747–11750 (1997).
10. Richmond, B. G. & Strait, D. S. *Nature* **404**, 382–385 (2000).
11. Dainton, M. *Nature* **410**, 324–325 (2001).
12. Arnason, U., Gullberg, A., Burguete, A. S. & Janke, A. *Hereditas* **133**, 217–228 (2000).
13. WoldeGabriel, G. *et al.* *Nature* **371**, 330–333 (1994).
14. Collard, M. & Wood, B. A. *Proc. Natl Acad. Sci. USA* **97**, 5003–5006 (2000).

ogy have made systematic ground-based searches much more tractable than even a decade ago. Today, digital CCD (charge-coupled device) arrays with 10,000 × 10,000 pixels can cover a quarter-square-degree patch of sky (roughly equivalent to the area of the full Moon) in a single exposure. Equally important are improvements in computers, which allow data to be analysed in real time, and the development of sophisticated algorithms that can automatically detect faint, slowly moving objects.

Despite these improvements, finding small satellites is still a daunting task. The region of space that needs to be covered in a thorough search is a planet's 'Hill sphere' — the area within which satellites can orbit stably. The projected area covered by the Hill sphere is determined by the planet's mass: for Jupiter, Saturn, Uranus and Neptune, these areas are 48, 22, 6 and 7 square degrees, respectively. Even with a quarter-square-degree field of view there is a lot of space to be covered and, partly for this reason, Gladman and colleagues focused first on more distant Uranus and Neptune^{2,3} before proceeding to Saturn¹. To cover Saturn's entire Hill sphere, Gladman *et al.* planned a careful systematic campaign involving multiple telescopes and coordinated follow-up observations.

Remarkably, this new survey covered Saturn's entire Hill sphere down to a brightness limit of 23rd magnitude¹. This means that, with reasonable assumptions for the reflectivities of the new moons, Gladman *et al.* have found all of the objects with radii larger than about 4 km that are circling Saturn (Fig. 1). They estimate that their previous studies of Uranus and Neptune are nearly complete, covering about 90% of the Hill sphere down to a similar brightness³. Although a complete survey of Jupiter's environs has not yet been undertaken, an impressive step in this direction was made earlier this year by Sheppard *et al.*⁴, who added ten new moons to the one detected last year⁵.

Because they shine by reflecting sunlight, satellites of equal brightness (23rd magnitude) around Jupiter, Saturn, Uranus and Neptune have very different radii: approximately 1, 4, 16 and 36 km, respectively. This makes it difficult to compare different satellite populations because the observations discriminate against the more distant planets. For example, all the new saturnian satellites are probably too small to have been detected if they had orbited Uranus or Neptune instead. Similarly, a group of three very faint objects that Gladman *et al.*¹ spotted moving near Saturn but then lost — when angling for moons, unlike when angling for fish, it is the small ones that get away — would have been more easily tracked had they circled Jupiter instead.

Nonetheless, complete surveys of satellite populations around individual planets can reveal important clues to the origin and

Planetary science

Saturn saturated with satellites

Douglas P. Hamilton

Advances in detector technology have led to a rash of newly discovered moons around the giant planets. Saturn currently has the most known satellites — but for how long?

Parents of small children are expected to know the answers to questions like "Which planets in the Solar System have the most moons?" But a surge of discoveries of distant planetary companions in the past five years has left beleaguered parents everywhere hard pressed to answer this query correctly. Today, the planets with the most known natural satellites are Saturn with 30 moons, Jupiter with 28, Uranus with 21 and Neptune with 8. But this list had a different order last year, and was different again the year before that. Not since the Voyager fly-bys of the 1980s have so many moons been discovered so quickly. On page 163 of this issue¹, an international team of astronomers led by Brett Gladman announces the

discovery of a dozen new kilometre-sized satellites — the culmination of their highly successful observations of Saturn.

The new discoveries have come not from telescopes in space, nor from the largest telescopes on Earth, but rather from medium-sized ground-based instruments (3–5 m diameter), which can efficiently scan large regions of the sky. Ground-based surveys are sensitive to small moons far from planets, but fail to spot nearby satellites owing to light scattered from the planet. They complement spacecraft measurements, which can easily find small moonlets close to their parent planet, but cannot efficiently search for distant satellites.

Rapid improvements in detector technol-

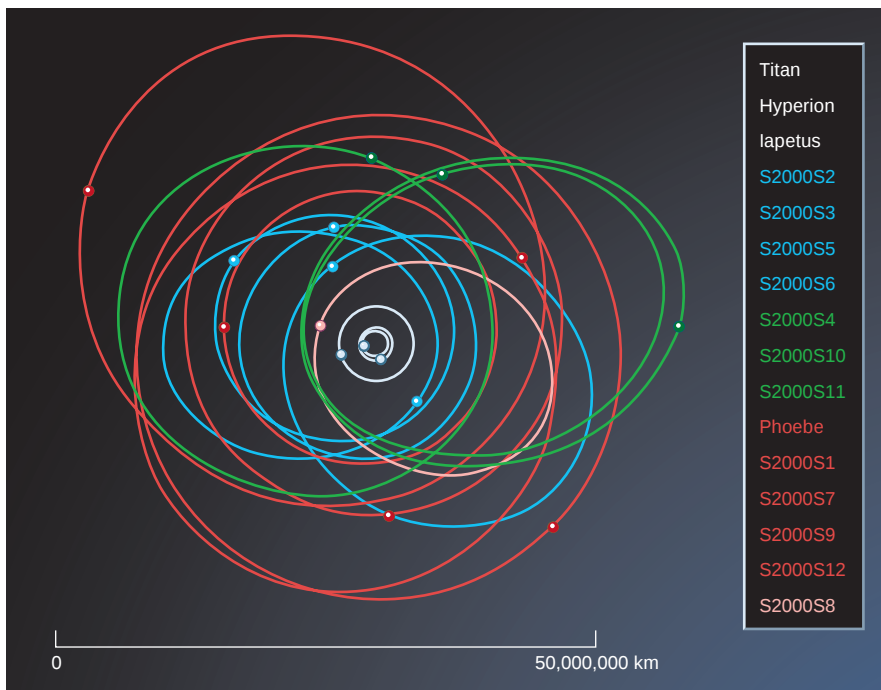


Figure 1 The current orbits of the outer satellites of Saturn. The radius of the Hill sphere is about 65 million km, which is equal to 1,100 Saturn radii, or 0.43 AU (1 AU is the mean Earth–Sun distance). Saturn is at the centre and the white objects on inner, nearly circular, prograde orbits are the classical satellites: Titan, Hyperion and Iapetus. Prograde objects circle Saturn in the same way as its classical satellites (anticlockwise when viewed from above), whereas retrograde objects orbit in the opposite direction (clockwise). The 12 new satellites discovered by Gladman *et al.*¹ can be grouped into families according to their orbital properties: cyan and green objects are the new prograde groups, whereas pink and red dots (including previously known Phoebe) are all retrograde satellites. Animations of these and all other Solar System satellites can be found at ref. 9.

early history of the planets. Indeed, Gladman *et al.* observe that Saturn's moons are grouped into distinct families (Fig. 1) with similar orbital properties, just like asteroids and distant satellites of Jupiter. Families are thought to be formed either during cratering collisions that break fragments off parent bodies, or during more violent catastrophic collisions that completely destroy the parent.

Although the current orbits of the satellites (Fig. 1) hint at these relationships, clearer indications come from similarities in their long-term orbital properties (particularly the average tilt of the orbital plane)¹. The existence of satellite families implies that there are substantial unseen populations of smaller bodies around all the giant planets, because collisions that chip 1-km fragments off 10-km moons occur far more frequently than collisions that produce 10-km fragments from 100-km moons. The three little ones that got away near Saturn are just the tip of the iceberg.

If each satellite family was derived from a single parent moon, where did the parents originate? It has long been believed that satellites were captured from independent orbits around the Sun early in the history of the Solar System. At that time, the Solar nebula, a large disk of gas and dust, still surrounded the Sun, and smaller disks circled each of the gaseous planets. There are several possible capture mechanisms involving interactions with the planet's gaseous nebula⁶, collisions with other objects within the planet's Hill sphere⁷, and expansion of the Hill sphere as the planet grows in size⁸. Although these mechanisms for capture have been investigated, the details are not well understood and the process of satellite capture remains an outstanding problem in planetary science.

The next observational challenge is a survey of Jupiter's entire Hill sphere down to 23rd magnitude. Such a survey is complicated by the large search area (more than for the other three giant planets combined), the potentially huge number of detectable 1-km-sized moonlets, the more rapid orbital motion of jovian satellites, and the greater likelihood of mistaking them for main-belt asteroids. Nonetheless, a multitude of jovian satellites probably lurk undetected in the vast region of space controlled by this giant planet. It will not be long before massive Jupiter regains

the title of the Solar System's most-mooned planet.

Douglas P. Hamilton is in the Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA.

e-mail: hamilton@astro.umd.edu

1. Gladman, B. J. *et al.* *Nature* **412**, 163–166 (2001).
2. Gladman, B. J. *et al.* *Nature* **392**, 897–899 (1998).
3. Gladman, B. J. *et al.* *Icarus* **147**, 320–324 (2000).
4. Sheppard, S. S. *et al.* *IAU Circ.* 7555 (2001).
5. Scotti, J. V. *et al.* *IAU Circ.* 7460 (2000).
6. Heppenheimer, T. A. & Porco, C. C. *Icarus* **30**, 385–401 (1977).
7. Pollack, J. B., Burns, J. A. & Tauber, M. E. *Icarus* **37**, 587–611 (1979).
8. Colombo, G. & Franklin, F. A. *Icarus* **30**, 385–401 (1971).
9. <http://janus.astro.umd.edu/solarsystem.html>

Apoptosis

Mostly dead

Douglas R. Green and Helen M. Beere

It has always been thought that once the process of cell suicide has passed a certain point, it is irreversible. Yet it seems that cells can recover — but only if they are not eaten by nearby 'phagocytic' cells.

In all animals, the process of programmed cell suicide (apoptosis) is coordinated by enzymes known as caspases, which cut up key substrates in the cell. The dying cell is then neatly packaged, engulfed by neighbouring 'phagocytic' cells, and cleared from the body without fanfare, leaving no evidence of the catastrophic events that preceded it. It has always been assumed that there is a 'point of no return' in this death cascade — at or shortly before the time at which caspases are activated — beyond which the process of cell execution proceeds inexorably. This view is challenged by Reddien *et al.*¹ and Hoepfner *et al.*² on pages 198 and 202 of this issue. It seems that cells in which

caspases have been activated can in fact progress through a state of being 'mostly dead', a stage that physically resembles the early phase of apoptosis but from which cells can fully recover.

During the development of the nematode worm *Caenorhabditis elegans*, 131 cells are destined to die by apoptosis. These cell deaths depend on the presence of a caspase, CED-3, and on a molecule called CED-4 that binds to and activates CED-3. In healthy cells, CED-4 is maintained in a functionally inactive state by its association with CED-9. The trigger for cell death is the protein EGL-1, which is expressed in response to developmental cues. EGL-1 binds to CED-9,

displacing CED-4, which in turn activates CED-3. The active CED-3 cleaves as-yet-unidentified substrates in the cell to promote apoptosis. In the absence of functional CED-3, no cell death occurs.

Several different mutations in the *ced-3* gene have been identified, the effects of which extend from partial to almost complete inhibition of cell death³. Using partial *ced-3*-mutant worms, Reddien *et al.*¹ and Hoepfner *et al.*² now show that cells that activate this caspase can begin the process of execution but then recover — but only if the cells are not engulfed by phagocytic cells. In *C. elegans*, seven genes, which form two partially redundant groups, are required for the engulfment and clearance of dying cells⁴. Both Reddien *et al.* and Hoepfner *et al.* observed that the combination of defects in phagocytosis with defects in caspase activity protected more cells from apoptotic death than did the caspase mutations alone. Moreover, they identified cells that recovered completely after undergoing morphological changes characteristic of early CED-3-dependent apoptosis. These 'mostly dead' cells were not found in animals completely lacking CED-3 activity. So it is reasonable to assume that CED-3 was active in the recovering cells and apoptosis began, but that the process then reversed. Caspase-dependent changes in a cell are not necessarily fatal.

How can apoptotic events be reversed? Unlike other signalling events, the cleavage of a substrate is irreversible. So two things must happen if a cell is to recover from caspase-induced morphological changes. First, there must be a mechanism to counteract or limit the activity of caspases. Good candidates are the inhibitor-of-apoptosis proteins (IAPs), some of which bind to and inactivate caspases in vertebrates and insects⁵. Although no such regulator has been identified in nematodes (with the exception of BIR-1, which seems to have no direct role in apoptosis⁶), our superficial search of the nematode genome reveals at least one candidate. A screen for mutations that increase cell death in CED-3-defective animals might reveal more.

Second, the effects of caspase-dependent cleavage events must be reversible. Caspases coordinate apoptosis through two general classes of substrate, which we have nicknamed 'swords' and 'shields' (Fig. 1). Swords are activated by caspases, and then catalyse events that contribute to cell death. Shields, which sustain normal cellular functions, are destroyed by caspases. If caspase-mediated execution of a cell can indeed begin and then be reversed, there must be mechanisms that inhibit or counteract the activated swords, and repair or replace damaged shields. These may be normal repair processes, but at this point we simply do not know.

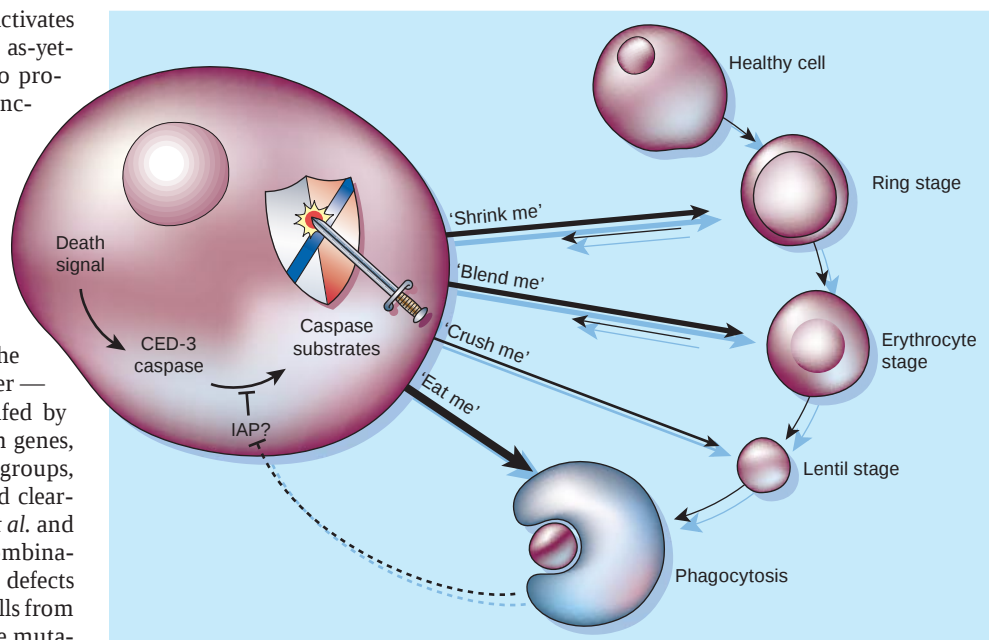


Figure 1 At the brink of death. In *Caenorhabditis elegans* cells that receive a death signal, the caspase enzyme CED-3 is activated and directs morphological changes in the dying cell. This occurs through cleavage of key substrates, which (although not yet identified in worms) are likely to fall into two camps: 'swords' and 'shields'. Swords are activated by cleavage and promote apoptosis; examples in mammals include ICAD–CAD, which cuts DNA. Shields support normal cellular functions until they are cleaved and inactivated; mammalian examples are poly-ADP-ribose polymerase and DNA-PK, both of which are involved in DNA repair. The substrates that produce the 'eat-me' signal are presumably the most active or most easily activated, as engulfment of apoptotic cells by phagocytes prevents or obscures recovery from the other stages^{1,2}. Phagocytes that bind to cells displaying 'eat-me' markers deliver a signal that enhances apoptosis. The signals that result in cytosolic shrinkage ('shrink me'), leading to the ring stage, or loss of nuclear–cytosol distinction ('blend me'), leading to the erythrocyte stage, are presumably reversible. The 'crush-me' events that lead to the lentil stage may not be reversible. IAP, inhibitor-of-apoptosis protein.

One example of a sword in vertebrates is the ICAD–CAD complex, which when cleaved by caspases becomes an active nuclease, breaking down nuclear DNA. Another nuclease, which is present in mammals and *C. elegans*, is released from mitochondria during apoptosis^{7,8}. The destruction of DNA should be irreversible, so we might assume either that it does not occur in the almost-dead cells, or that these cells can persist and differentiate despite such irreparable damage. Perhaps different caspase substrates vary in their susceptibility to caspases; this hierarchy of susceptibility might govern the point at which the cell is irreversibly committed to death. Or maybe only one CED-3-mediated cleavage event is of any consequence — that which reveals an 'eat-me' signal to phagocytes. If so, in the absence of effective phagocytosis, the cell would be able to recover.

The authors^{1,2} also conclude that phagocytic cells actively promote cell death, although this proposed effect does not occur when CED-3 is inactive. The simplest explanation here is that when phagocytosis begins — which happens only after the dying cell has generated the CED-3-dependent 'eat-me' signal — digestive enzymes from the phagocyte are released before engulfment

occurs. Alternatively, the phagocytes might somehow overcome the apoptosis-blocking signals from IAPs (for example). In other words, phagocytes might provide signals that accelerate apoptosis.

Those of us who are 'backbone-centric' might be tempted to discount these observations as having no relevance to vertebrates. Developmental cell death in the worm depends completely on CED-4 and CED-3. But, in mammalian cells, the inhibition of caspases or knockout of the mammalian CED-3 or CED-4 counterparts prevents apoptosis but not cell death — the cells die more slowly and by a different process⁹. This cell death might relate to loss of mitochondrial function after permeabilization of the outer mitochondrial membrane, an event that is required for an important apoptosis pathway in vertebrates but probably not in nematodes¹⁰.

That said, mice that lack these proteins do have developmental defects associated with reduced apoptosis^{11,12}. And cells from these animals resist the apoptosis that guards against cancer, and so are more susceptible to transformation than normal cells¹¹. As in worms, these knockouts affect processes that occur after the presumed point of no return. So there may be mechanisms in

mammalian cells that reverse the process of apoptosis after this point. Perhaps vertebrate cells can also be 'mostly dead' at some time during their execution.

In nematodes, the failure to remove apoptotic cells does not seem to have harmful consequences (although cell corpses accumulate). But mice that lack the tyrosine kinase MER have defects in clearing apoptotic cells by phagocytosis, and this promotes autoimmune disease¹³. Arguably, then, the enhancement of death by phagocytes may ensure the coordination of death and clearance. This would reduce the risk of any cellular contents leaking from the dying cell and initiating an inflammatory response. And phagocyte-enhanced cell death might also help to control cancer in vertebrates. In mice, neutrophil cells that are rendered resistant to apoptosis can nevertheless be eliminated by phagocytosis¹⁴; if combined with a phagocytosis defect, these cells might survive and accumulate. It will be interesting to see whether the defect in clearing apoptotic cells in MER-deficient mice can promote cancer, especially when combined with defects in caspase activation.

In William Goldman's screenplay of

The Princess Bride, Miracle Max informs us that "there's a big difference between mostly dead and all dead... mostly dead is slightly alive". Apparently this applies to cells as well. ■

Douglas R. Green and Helen M. Beere are at the La Jolla Institute for Allergy and Immunology, 10355 Science Center Drive, San Diego, California 92121, USA.

e-mail: dgreen5240@aol.com

1. Reddien, P. W., Cameron, S. & Horvitz, H. R. *Nature* **412**, 198–202 (2001).
2. Hoepfner, D. J., Hengartner, M. O. & Schnabel, R. *Nature* **412**, 202–206 (2001).
3. Shaham, S., Reddien, P. W., Davies, B. & Horvitz, H. R. *Genetics* **153**, 1655–1671 (1999).
4. Gumieny, T. L. & Hengartner, M. *Cell Death Differ.* (in the press).
5. Deveraux, Q. L. & Reed, J. C. *Genes Dev.* **13**, 239–252 (1999).
6. Speliotes, E. K., Uren, A., Vaux, D. & Horvitz, H. R. *Mol. Cell* **6**, 211–223 (2000).
7. Parrish, J. et al. *Nature* **412**, 90–94 (2001).
8. Li, L. Y., Luo, X. & Wang, X. *Nature* **412**, 95–99 (2001).
9. Chautan, M., Chazal, G., Cecconi, F., Gruss, P. & Golstein, P. *Curr. Biol.* **9**, 967–970 (1999).
10. Meier, P., Finch, A. & Evan, G. *Nature* **407**, 796–801 (2000).
11. Kuida, K. et al. *Cell* **94**, 325–337 (1998).
12. Cecconi, F., Alvarez-Bolado, G., Meyer, B. I., Roth, K. A. & Gruss, P. *Cell* **94**, 727–737 (1998).
13. Scott, R. S. et al. *Nature* **411**, 207–211 (2001).
14. Lagasse, E. & Weissman, I. L. *J. Exp. Med.* **179**, 1047–1052 (1994).

Nanotechnology

Less is more

J. Tersoff

Creating a structure as simple as a hole can be a challenge — when the hole is just a few nanometres wide. The trick is to start small and then get smaller.

In the quest for nanoscale technologies, smaller is better. Even the humble hole, if small enough, has many possible applications, ranging from molecular detectors to masks for stencils. In cells, such nanoscale pores play a crucial role in controlling the transport of ions and molecules through the cell membrane. Now a group of physicists, engineers and biologists at Harvard¹ (writing on page 166 of this issue) has fabricated a 'nanopore', a hole just a few nanometres in diameter, in a thin but tough silicon nitride membrane. They demonstrate the potential importance of such a structure by using it to detect single molecules of DNA in solution.

Nanopores have previously been fabricated in organic materials, by incorporating channel-forming peptides into a lipid membrane². But many contexts where nanopores could be useful require hard, stable inorganic materials. For these materials, established techniques such as ion-beam cutting cannot currently fabricate structures just a few nanometres across. As an alternative, there has been intense interest in 'self-assembly', whereby nanoscale structures assemble themselves³ from featureless starting mater-

ials, through forces intrinsic to the system. In fact, arrays of nanoscale pores as small as 50 nm in diameter have been fabricated in this way⁴. This approach is generally restricted to large arrays, however, and it has not yet proven possible to generate structures of only a few nanometres in this way. To place a single molecular-scale structure in a selected location requires a different approach.

The Harvard group¹ refer to their approach as 'ion beam sculpting', in analogy with the way a chisel removes material to shape a marble sculpture. The original goal was to make a nanopore by simple ion-beam cutting, in which a beam of ions knocks atoms from the surface of a solid. But the new method relies on their serendipitous discovery that the beam can, in effect, add material as well as remove it. At first, the ion beam failed to drill any hole at all. Further investigation revealed that the beam would actually cause a pre-existing hole to shrink, and ultimately to close up. Through a series of experiments, along with theoretical modelling, the authors developed a convincing explanation for this surprising behaviour.

Ion beams are widely used in materials

processing, but their effects are generally too complex to work out in detail. Here, by studying the delicate balance between the opening and closing of the hole, the authors gain new insight into the interaction between ion beam and solid. As well as removing atoms, the ion beam knocks some atoms loose, so they can diffuse along the surface. Diffusing atoms are drawn into the pore from the nearby region, perhaps by capillary forces, causing the hole to shrink. These cutting and healing effects compete; which one dominates depends on the beam intensity and the temperature. Aside from any practical applications of nanopores, these results provide an intriguing window onto the effect of ion beams on materials.

To form extremely small pores, one begins with a larger hole (say, 60 nm in diameter), and exposes it to the ion beam at a temperature and beam intensity for which the healing effect dominates. The ions passing through the pore are counted, providing a measure of the pore size. The ion beam is switched off when the desired pore size is achieved. (If the process is allowed to proceed unchecked, the pore will close up completely.) The method is similar to self-assembly in that a structure is created that is smaller than any dimension over which one has direct control. Instead, the new approach relies on a feedback system that allows the scientists to monitor the process and stop it at the desired time. Such a strategy has also been used to fabricate nanoscale electrical contacts⁵ that are smaller than it is possible to make with any direct method. But in that case, rather than a hole, the nanoscale structure is a metallic path for electrons, formed by the growth of an oxide bridge across a larger opening. This feedback approach may prove generally attractive for extending the limits of nanoscale fabrication.

The authors illustrate the potential utility of the nanopore by using it as a simple electronic detector of single DNA molecules. A silicon nitride membrane containing one nanopore is placed between two compartments of saline solution, separating them electrically. A small voltage (120 millivolts) is applied across electrodes in the two regions, providing a tiny ionic current (about 2 nanoamperes) through the pore. When double-stranded DNA is added to the solution on the negative side, the current is blocked intermittently, presumably as the negatively charged molecules are drawn towards the positive side and partly block the pore. Similar detectors have previously been made using organic materials²; but silicon nitride has the advantage of being mechanically and chemically robust, and allows the fabrication of rigid pores of any chosen size.

The authors suggest that their technique could be used to fabricate structures other than nanopores, such as thin slits or trenches, and also using other materials including

aluminum, silicon and silicon dioxide. They also suggest that, beyond making a single nanostructure, improvements in ion-beam technology will allow highly parallel fabrication of nanostructures. If these ideas work out in practice, the method could find many applications in scientific studies and, ultimately, in new technologies. ■

J. Tersoff is at the IBM Thomas J. Watson Research

Center, PO Box 218, Yorktown Heights, New York 10598, USA.

e-mail: tersoff@us.ibm.com

1. Li, J. et al. *Nature* **412**, 166–169 (2001).
2. Bezrukav, S. M., Vodyanoy, I. & Parsegian, V. A. *Nature* **370**, 279–281 (1994).
3. Ball, P. *Nature* **379**, 25 (1996).
4. Li, A. P., Müller, F., Birner, A., Nielsch, K. & Gösele, U. *J. Appl. Phys.* **84**, 6023–6026 (1998).
5. Schmidt, T., Martel, R., Sandstrom, R. L. & Avouris, P. *Appl. Phys. Lett.* **73**, 2173–2175 (1998).

Developmental biology

Vesicles and the spinal cord

Juhee Jeong and Andrew P. McMahon

The distinction between cell biology and developmental biology is becoming increasingly blurred. The latest example involves a signalling pathway switched on in the developing spinal cord.

Membrane-clad containers known as vesicles are used to distribute proteins and other molecules to appropriate locations within a cell, and so are essential for cells (and organisms) to function normally. On page 194 of this issue¹, Eggenschwiler and colleagues propose an intriguing new function for vesicle trafficking. They reveal a previously unsuspected link between a protein implicated in vesicle transport and a signalling pathway with a key role in the formation of different types of neuron in the developing spinal cord.

Proteins of the Hedgehog family are secreted signalling molecules that are involved in intercellular communication.

They are essential for several processes in the development of invertebrate and vertebrate embryos². For example, Sonic hedgehog (Shh) is required for patterning the mammalian spinal cord, inducing distinct ventral neurons at specific positions in the developing neural tube in a concentration-dependent mechanism³. So, for instance, floor-plate cells develop closest to the source of Shh in response to the highest concentration of Shh, and V3 interneurons at a more distant position in response to a lower concentration. The incorrect regulation of the Hedgehog signalling pathway has also been implicated in several cancers². Yet, despite their obvious importance, the mechanisms

that regulate Hedgehog signalling are only just being uncovered.

In the fruitfly *Drosophila*, the activation of Hedgehog-responsive pathways depends on interactions between two integral membrane proteins: Patched, which contains a region that might sense sterols (components of cell membranes), and Smoothened, a protein with similarity to G-protein-coupled receptors⁴ (a broad class of signalling molecules). In the absence of Hedgehog, the task of Patched is to inhibit Smoothened, thereby blocking the whole Hedgehog-responsive pathway. When bound to Hedgehog, Patched can no longer block Smoothened, so the pathway is activated (one of the outcomes of this is the increased expression of Patched itself). This basic mechanism appears to be conserved from flies to humans. What remains unclear is exactly how Patched inhibits Smoothened, and how this inhibition is eliminated when Hedgehog binds to Patched.

Early studies suggested a direct interaction between Patched and Smoothened⁵. Although this may indeed be the case for a small amount of each protein, there is little overlap in the intracellular distributions of these proteins in cells that detect the Hedgehog signal^{6,7}. So an indirect mechanism is now invoked, based on the observation that the binding of Hedgehog to Patched results in Smoothened becoming more highly phosphorylated and more stable^{6,8}. These changes correlate with the accumulation of Smoothened at the cell surface, and with the internalization and destabilization of

Evolutionary biology

Autumn colour code

Most of us are happy simply to marvel at the autumn colours of deciduous trees, as shown here. But the late William Hamilton, one of the most influential thinkers on evolution in the twentieth century, and Sam Brown felt compelled to ask what such a show is in aid of. They hypothesize that trees are sending the message 'pick on someone else' to their insect enemies.

Most researchers, when they have considered autumn displays at all, have assumed them to be a side effect of senescence. But Hamilton and Brown felt that trees would not, without good reason, make large amounts of potentially costly compounds in their leaves just before shedding them.

In autumn, many insects are looking for plants on which to

overwinter and to feed and reproduce the next year. Hamilton and Brown concentrated on aphids, because aphid species tend to be choosy about what they eat and use colour cues to find their hosts. The amount of damage aphids can inflict also gives trees a strong incentive to deter them.

Using published data, Hamilton and Brown surveyed 262 tree species and found that the yellowness or redness of a tree's autumn leaves correlates with the number of aphid species that attack it (*Proc. R. Soc. Lond. B* **268**, 1489–1493; 2001). Maples, for example, which put on some of the most spectacular displays, are some of the most heavily aphid-infested species, fitting in with the idea that tree species suffering greater insect damage



should invest more in colour signalling. Costly autumn pigments would be an 'honest signal' — only trees that were truly committed to defence would make them.

The connection between colours and herbivores raises questions for

exploring the hypothesis further. For instance, does the link hold good within species, as well as between them — that is, are brighter individuals left alone? And can the idea be broadened beyond aphids? **John Whitfield**

RICHARD CUMMINS/CORBIS

Daedalus

Bashing the bugs

Britain's National Health Service has been accused of giving patients diseases they didn't have when they entered hospital. Indeed, a recent survey claims that a third of British NHS wards fall below basic sanitary standards. And one study found that certain multiple-antibiotic-resistant bacteria, such as MRSA (methicillin-resistant *Staphylococcus aureus*) abound on specific wards. In this connection, Daedalus recalls an old American remedy.

Late nineteenth-century warehouses in the United States were often infested with rats, cockroaches and bugs of all sorts. The cure was to close all the windows, and place in every room a large pellet of sodium cyanide, together with an apparatus for covering this with hot sulphuric acid. When all was ready, the operator left the building, pulling the strings which worked the apparatus in each room. The whole warehouse became a gas chamber, killing all the infesting organisms. A little while later, the windows were opened (from the outside!). Later still, the building, now free of pests, could be safely reoccupied by its human owners.

So, says Daedalus, let us adapt this technology for the NHS. Hydrogen cyanide should be readily available in cylinders or adsorbed on vermiculite, for example; other volatile agents to kill viruses, such as formaldehyde or ethylene oxide, are also freely accessible. (Formaldehyde has already been blamed for affecting people in houses whose cavity walls have been filled with urea-formaldehyde waterproof resin.) Mixtures of ethylene oxide and carbon dioxide, just as deadly to infestations but less lethal to human beings, have also been used in pest control.

The NHS patients would have to be rehoused in a day-room or vacant ward, and external fans might have to be installed to prevent gas issuing from the treated ward from inconveniencing surrounding wards or dwellings. But what worked for early Americans should also work well for the NHS. Insects, bacteria and viruses should be neatly eliminated, and patients would be safely rehoused in completely sterile surroundings.

The only snag, says Daedalus, might be the bugs left in the patients themselves. The American solution was to disinfect the warehouses at regular intervals, so that nothing nasty could build up even by mutation. Whether NHS hospitals could build such a procedure into their routine is another matter. But if patients arrive free of these infections, the problem should not arise.

David Jones

Patched. One model put forward to explain these data is that the phosphorylation of Smoothened is crucial to the signalling process, and that, in the absence of Hedgehog, Patched inhibits Smoothened by regulating — directly or indirectly — a dephosphorylating enzyme (phosphatase)⁶.

More recently, mutations have been detected in the sterol-sensing domain of Patched that render it unable to repress Smoothened, but do not affect its binding to Hedgehog^{7,9}. Together with the fact that other proteins with sterol-sensing domains, such as SCAP and NPC1, are thought to control vesicle transport, this hints that Patched might also regulate vesicle movement.

The idea that vesicle transport is in some way involved in Hedgehog signalling now receives support from Eggenchwiler *et al.*'s studies¹ of mice with mutations in the *open-brain* (*opb*) gene. Such mutations lead to the neural tube forming improperly, being open in both the brain and spinal cord. Two independently arising mutations have been isolated in *opb*^{10,11}, and embryos with these mutations have defects that are characteristic of overactive Shh signalling — for example, suppression of dorsal and overproduction of ventral neuronal precursors^{10,12}.

Eggenchwiler *et al.*¹ show that mice with mutations in both *opb* and *Shh* have similar defects to animals with mutations in *opb* alone. Remarkably, despite the absence of a localized source of Shh, the double mutants have a reasonably normal positioning of floor-plate cells and V3 interneurons in the ventral neural tube. The data suggest that a gradient of Shh is not absolutely essential either for establishing the polarity of ventral cells or for the correct graded expression of the mouse *Patched1* gene. This is interesting because the expression of *Patched1* directly reflects the strength of signalling through the Shh pathway. Shh is clearly not present in these mice, so what could be going on here?

Part of the answer is probably that the protein product of the *opb* gene somehow inhibits signalling through the Shh pathway. Mutation of *opb* results in Shh-independent activation of the pathway. As the authors point out, bone morphogenetic proteins (which cause cells to take on dorsal identities³) and their antagonist Noggin¹³ may provide another way to establish ventral patterning.

What, then, is the protein encoded by the *opb* gene? Positional cloning provided Eggenchwiler *et al.* with the answer¹: *opb* encodes Rab23, a member of a large family of GTP-hydrolysing enzymes (GTPases). Although the functions of Rab23 in particular have not been investigated, Rab proteins in general are master regulators of vesicle trafficking. By serving as a scaffold for other molecules, Rab proteins coordinate the budding of vesicles from one cellular compart-

ment, their transport, and their docking and fusion with the target compartments¹⁴.

The implication is that vesicle trafficking is important in the Hedgehog signalling pathways. Moreover, as the detection of the Hedgehog signal by *Drosophila* cells leads to an alteration in the localization of Smoothened and Patched, it is tempting to speculate that Rab23 participates in a vesicle-transport process that promotes the inhibition of Smoothened by Patched1 in mice. For example, if the phosphorylation of Smoothened is indeed important for its activity, then Patched1 might, in a Rab23-dependent way, direct vesicles containing Smoothened (or a phosphatase that acts on Smoothened) to a cellular compartment in which Smoothened is dephosphorylated or destabilized.

Eggenchwiler *et al.*'s work¹ offers several testable predictions. First, if Rab23 is indeed required for the inhibition of Smoothened by Patched1, then ventral cell identities should be lost in mice with mutations in both *opb* and *smoothened*, much as in *Shh* mutants. Second, Smoothened or Patched1 might be localized incorrectly in *opb* mutant cells. Third, the defects in mice with mutations in *opb* alone are less severe than in mice lacking *Patched1* (ref. 15), so the inhibition of Smoothened might not be completely blocked in *opb* mutants. Reducing the levels of Patched1 in *opb* mutants might then be expected to enhance the defects. Finally, given that the sterol-sensing domain of the fruitfly Patched protein is essential for the inhibition of Smoothened, this domain might be needed in some way for the Rab23-mediated regulation of vesicle transport. If so, Rab23 might also control vesicle movement by working with other proteins that contain a sterol-sensing domain. ■

Juhee Jeong and Andrew P. McMahon are in the Department of Molecular and Cellular Biology, The Biolabs, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

e-mail: amcmahon@mcb.harvard.edu

1. Eggenchwiler, J. T., Espinoza, E. & Anderson, K. V. *Nature* **412**, 194–198 (2001).
2. Hammerschmidt, M., Brook, A. & McMahon, A. P. *Trends Genet.* **13**, 14–21 (1997).
3. Jessell, T. M. *Nature Rev. Genet.* **1**, 20–29 (2000).
4. Ingham, P. W. *EMBO J.* **17**, 3505–3511 (1998).
5. Stone, D. M. *et al.* *Nature* **384**, 129–134 (1996).
6. Deneef, N., Neubüser, D., Perez, L. & Cohen, S. M. *Cell* **102**, 521–531 (2000).
7. Martin, V., Carrillo, G., Torroja, C. & Guerrero, I. *Curr. Biol.* **11**, 601–607 (2001).
8. Kalderon, D. *Cell* **103**, 371–374 (2000).
9. Strutt, H. *et al.* *Curr. Biol.* **11**, 608–613 (2001).
10. Günther, T., Struwe, M., Aguzzi, A. & Schughart, K. *Development* **120**, 3119–3130 (1994).
11. Kasarskis, A., Manova, K. & Anderson, K. V. *Proc. Natl Acad. Sci. USA* **95**, 7485–7490 (1998).
12. Eggenchwiler, J. T. & Anderson, K. V. *Dev. Biol.* **227**, 648–660 (2000).
13. McMahon, J. A. *et al.* *Genes Dev.* **12**, 1438–1452 (1998).
14. Rodman, J. S. & Wandinger-Ness, A. J. *Cell Sci.* **113**, 183–192 (2000).
15. Goodrich, L. V., Milenkovic, L., Higgins, K. M. & Scott, M. P. *Science* **277**, 1109–1113 (1997).

Obituary

Rosa Beddington (1956–2001)

Rosa Beddington's death on 18 May cut short her career at the untimely age of 45, but not before she had rewritten the textbooks. Generations of biologists had been taught about the classical studies that had identified the Spemann organizer, a region in amphibian embryos that emits signals responsible for inducing adjacent tissues to form the body axis, complete with head and tail. Working in the mouse, Beddington made the greatest contribution to understanding axis formation since those original 1920s observations. Her experiments revealed that the anterior–posterior axis of the vertebrate embryo forms as the result of a cooperative interaction between two sources of inducing activities that we now term the 'head' and 'trunk' organizers. Perhaps more notably, her studies helped to uncover the importance of the mammalian extra-embryonic tissues for induction of the future head; these tissues had previously been viewed as mere evolutionary add-ons required for viviparity.

As a D.Phil. student at the University of Oxford in the late 1970s, Beddington took on the challenge of understanding events in the early-stage mouse embryo. Mouse embryos grow outside the protective environment of the uterus only for short periods of time. Moreover, early axis formation is accompanied by extraordinary growth and cell movements. Undaunted by these complexities, Beddington painstakingly grafted pieces of tritium-labelled tissue from different regions of donor embryos to various sites in a host embryo. She then scrutinized histological sections to see whether donor tissues could adopt new fates or whether the graft's developmental potential was restricted by its site of origin. Together with work by Kirstie Lawson and Patrick Tam, her papers provided the all-important fate map of post-implantation lineages in the mammalian embryo. Without this basic information, all the subsequent work on how, when and what a cell is instructed to become in the embryo would have been impossible.

At the Imperial Cancer Research Fund laboratories in Oxford, Beddington harnessed the new transgenic technology to generate strains of mice expressing β -galactosidase. She fondly dubbed her most famous animal Levi — a mouse with 'blue genes' that simplified analysis of tissue grafts and embryo manipulation. She also started working on embryonic stem cells, made from mutant mice, that provided



Exceptional talent and influence in developmental biology

insights into the cellular defects underlying the truncated body axis of the classical T (brachyury) mutation.

On moving in 1991 to the Centre for Genome Research in Edinburgh, she decided that the future of mouse embryology lay in understanding the embryo's own molecular mechanisms. Together with Bill Skarnes, she developed gene-trap vectors to identify the molecules expressed in different regions of the embryo. She also made complementary DNA libraries from tissue she herself dissected, to obtain cells corresponding to the germ layers (endoderm, mesoderm and ectoderm) of the early embryo. These libraries were widely distributed and proved a rich source of new genes.

In 1993, during a period of great progress in understanding the signals responsible for mesoderm formation in frog embryos, Beddington moved to the National Institute for Medical Research at Mill Hill in north London, to set up a new division of mammalian embryology. Convinced that all vertebrate embryos had developmental pathways in common, she performed grafting experiments with a mouse node, the 'classical organizer'. These were a technical *tour de force* but, contrary to findings in the frog, the result was only partial axis duplication. She worried about this and ultimately discovered the basis for the discrepancy. Analysis of gene-expression patterns led her to observe that a handful of genes were expressed asymmetrically in the extra-embryonic tissue of the visceral endoderm, long before any visible signs of the axis became apparent. In particular, she showed

that certain transcription factors are expressed on the prospective anterior side of the embryo many hours before the onset of gastrulation, several days before the node forms. She discovered that the initial proximal–distal polarity of the embryo is converted to the anterior–posterior axis by a coordinated movement of the visceral endoderm to the prospective anterior side. Thus she identified the anterior visceral endoderm as the signalling centre responsible for anterior patterning during early mammalian development. She was elected a fellow of the Royal Society in recognition of this seminal work. With her Mill Hill colleagues, she showed that similar processes occur in frog development.

For many years Beddington helped to teach the Cold Spring Harbor course on mouse molecular embryology. In teaching, too, she was exceptional. With her exquisite hand-drawn illustrations, she effortlessly conveyed the complexity of the mouse embryo, and often improvised using a felt pen and polystyrene coffee cup to explain gastrulation. Like all great biologists, she had a breadth of vision and a true understanding of all vertebrates, allowing her insights denied to others. A fierce critic, as tough on herself as on others, she did not tolerate fools gladly and liked arrogance even less.

Rosa was refreshingly modest for someone of her prodigious talents and intellectual range, often expressing doubts as to the significance and quality of her own work. She was always critical of her contemporaries — and certainly didn't cut us any slack — yet was exceptionally warm and generous with junior colleagues, in turn engendering fierce loyalty from them. In dying, as in living, she defied expectations, extending predictions of a few weeks' life expectancy into almost three years.

She will be remembered in good times as we learn more about the embryo and miss knowing what she would think about our experiments, and perhaps even more in the difficult periods when she would inevitably insist we should rise to the occasion and battle on, no matter what the obstacles. Rosa had a profound influence on her chosen field of study, an achievement she allowed herself. She was perhaps less aware of the lasting impression she made on all of us who shared her life.

Sohaila Rastan and Elizabeth Robertson
Sohaila Rastan is at Ceros Ltd, Wellington House, East Road, Cambridge CB1 1BH, UK. Elizabeth Robertson is in the Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mails: sohaila@rastan.fsnet.co.uk
ejrobert@fas.harvard.edu

Sex-biased dispersal of great white sharks

In some respects, these sharks behave more like whales and dolphins than other fish.

Ecological information about populations of great white sharks (*Carcharodon carcharias*) has been difficult to acquire, not least because of the rarity and huge size of this fish. Here we use genetic methods to show that the dispersal of *C. carcharias* is sex-biased, with philopatric (non-roving) females and roving males. In conjunction with other shared life-history features, including low fecundity, long lifespan and late age at maturity, our findings indicate that the population biology of *C. carcharias* may be more similar to that of marine mammals^{1,2} than to that of other fish.

Carcharodon carcharias (Fig. 1) is globally distributed in temperate waters off continental shelves³ and investigations have been undertaken on several populations^{4,5}. The analysis described here focuses on sharks collected off the coasts of South Africa (SA), Australia (AU) and New Zealand (NZ), where we were able to sample a relatively large number of individuals.

Pairwise comparison of complete

sequences from the non-coding 'control' region of the maternally inherited mitochondrial genome reveals two divergent genetic lineages within *C. carcharias* which differ by about 4% in sequence (Fig. 2a). Estimates of population differentiation on the basis of the control-region sequences (using F_{ST} analysis) show that *C. carcharias* from AU and NZ coastal waters are not significantly different, but that individuals from these populations are distinct from SA sharks (F_{ST} values of 0.81 and 0.89 for pairwise estimates between SA and AU, and SA and NZ, respectively; $P < 0.0001$ for each). The estimated differentiation approaches the theoretical maximum of unity, which suggests long-term isolation of these populations.

A survey of restriction-fragment-length polymorphisms in mitochondrial DNA from further individuals (Fig. 2b) supported this result, with only 1 of the 95 animals surveyed found to be out of place — a 3.5-metre male captured in Tasmania (AU) with a South African haplotype. These results suggest that female-mediated gene flow between ocean basins is rare.

In contrast, population genetic analysis of the sharks from AU and NZ ($n = 52$) and SA ($n = 43$) for five polymorphic nuclear-encoded microsatellite loci⁶ reveals no significant differentiation (Table 1). The low differentiation between populations that was identified by using nuclear DNA markers suggests that male-mediated gene flow occurs at a level that is sufficient to homogenize allele frequencies, and that dispersal of individuals is more extensive than has been indicated by movements estimated from tagging data⁵.

The contrast between the sequence differentiation revealed for the maternally inherited genetic marker and the lack of nuclear-gene differentiation indicates that female sharks are probably philopatric and that males may undertake transoceanic excursions. Although we have no direct



Figure 1 The great white shark (*Carcharodon carcharias*), which, although rare, inhabits temperate waters throughout the world.

evidence for this idea, sex-specific differences in habitat use in several shark species^{7,8}, including *C. carcharias*^{5,9}, and in migratory behaviour¹⁰ have been reported. These and other similarities between sharks and marine mammals suggest that general theories of sex-biased dispersal in birds and mammals¹¹ may also be relevant to *C. carcharias* and other species of shark.

Carcharodon carcharias is categorized as 'vulnerable' on the World Conservation Union's 'red' list of threatened animals³, so our findings may have implications for conservation strategies. Exploitation of a population in which immigration from surrounding stocks may be low and females are philopatric could lead to a rapid decline in stock size and future sustainability. The global dispersion of males may indicate that the demography of widely separated populations is linked, underscoring the need for international regulations to govern exploitation of great white sharks.

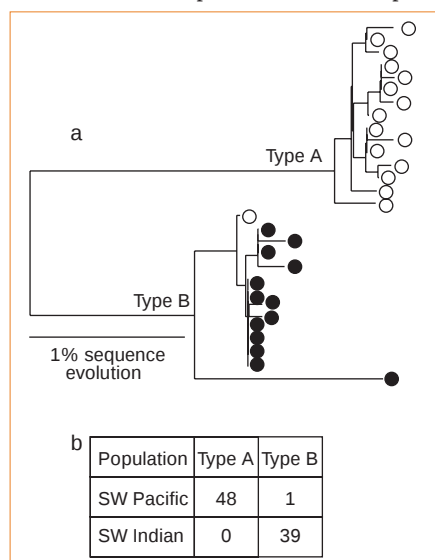


Figure 2 Sequence and survey of restriction-fragment-length polymorphisms (RFLPs) of the mitochondrial-DNA control region from great white shark populations reveals two divergent mitochondrial DNA types (called A and B). **a**, Phylogram depicting the relationships between individuals for which complete control-region sequences were determined. Black circles, samples from South African population; white circles, samples from Australian/New Zealand populations. (GenBank accession numbers for the D-loop sequences are AY026196–AY026224). **b**, Summary of the distribution of haplotypes determined using RFLP analysis of control-region sequences amplified by polymerase chain reaction. The observed distribution is significantly different from random expectations; Fisher's exact test, 0.958; $P < 0.0001$. Full details of methods and further information are available at <http://stripe.colorado.edu/~am/WhiteSharks.html>

Table 1 Nuclear-encoded microsatellite data from great white sharks

Locus*	No. of alleles			Observed heterozygosity		$\theta^\dagger$	$\rho^\ddagger$
	AU/NZ	SA	All	AU/NZ	SA		
<i>Ccar1</i>	6	5	7	0.686	0.643	NS	NS
<i>Ccar6</i>	4	5	5	0.647	0.595	NS	NS
<i>Ccar9</i>	15	12	15	0.882	0.857	NS	NS
<i>Ccar13</i>	9	7	9	0.804	0.786	NS	NS
<i>Ccar19</i>	4	3	4	0.451	0.452	NS	NS
All	38	32	40	0.694	0.667	NS	NS

Two populations of great white sharks were surveyed: Australian–New Zealand (AU/NZ) and South African (SA). The results indicate that there is a lack of significant differentiation between the two populations. NS, not significant. Further information is available at:

<http://stripe.colorado.edu/~am/WhiteSharks.html>

*GenBank accession number is AF184085 for microsatellite *Ccar6*; for information about the other loci, see ref. 6.

$\dagger$ This statistic is an analogue of F_{ST} (ref. 12), a measure of population differentiation.

$\ddagger$ Unbiased estimator of Slatkin's R_{ST} (ref. 13), another measure of population differentiation.

Amanda T. Pardini*, Catherine S. Jones*,
Leslie R. Noble*, Brian Kreiser†‡,
Hamish Malcolm§, Barry D. Bruce§,
John D. Stevens§, Jeremy Cliff||,
Michael C. Scholl¶, Malcolm Francis#,
Clinton A.J. Duffy☆, Andrew P. Martin‡

*Department of Zoology, University of Aberdeen,
Tillydrone Avenue, Aberdeen AB24 2TZ, UK

†Department of Biological Sciences, University of
Southern Mississippi, Hattiesburg,
Mississippi 39406, USA

‡Department of Environmental, Population and
Organismic Biology, University of Colorado,
Boulder, Colorado 80309, USA

e-mail: am@stripe.colorado.edu

§CSIRO Marine Research Laboratories, Castray
Esplanade, Hobart, Tasmania 7000, Australia

||Natal Sharks Board, Private Bag 2,
Umhlanga Rocks 4320, South Africa

¶PO Box 1258, Gansbaai 7220, Western Cape,
South Africa

#National Institute of Water and Atmospheric
Research, PO Box 14-901, Kilbirnie, Wellington,
New Zealand

☆Scientific & Research Unit, Department of
Conservation, PO Box 112, Hamilton, New Zealand

- Escoriza-Trevino, S. & Dizon, A. E. *Mol. Ecol.* **9**, 1049–1060 (2000).
- Lyrholm, T., Leimar, O., Johanneson, B. & Gyllenstein, U. *Proc. R. Soc. Lond. B* **266**, 347–354 (1999).
- Compagno, L. J. V., Marks, M. A. & Fergusson, I. K. *Environ. Biol. Fish.* **50**, 61–62 (1997).
- Strong, W. R. Jr, Bruce, B. D., Nelson, D. R. & Murphy, R. D. in *Great White Sharks: the Biology of Carcharodon carcharias* (eds Klimley, A. P. & Ainley, D. G.) 401–414 (Academic, San Diego, 1996).
- Cliff, G., Van der Elst, R. P., Govender, A., Witthuhn, T. K. & Bullen, E. M. in *Great White Sharks: the Biology of Carcharodon carcharias* (eds Klimley, A. P. & Ainley, D. G.) 393–400 (Academic, San Diego, 1996).
- Pardini, A. T., Jones, C. S., Scholl, M. C. & Noble, L. R. *Mol. Ecol.* **9**, 1171–1193 (2000).
- Economakis, A. E. & Lobel, P. S. *Environ. Biol. Fish.* **51**, 129–139 (1998).
- Klimley, A. P. *Environ. Biol. Fish.* **18**, 27–40 (1986).
- Klimley, A. P. *South. Calif. Acad. Sci. Mem.* **9**, 15–40 (1985).
- Bass, J. A. in *Sensory Biology of Sharks, Skates and Rays* (eds Hodgson, E. S. & Mathewson, R. F.) 545–594 (Office of Naval Research, Arlington, Virginia, 1978).
- Greenwood, P. J. *Anim. Behav.* **28**, 1140–1162 (1980).
- Weir, B. S. & Cockerham, C. C. *Evolution* **38**, 1358–1370 (1984).
- Goodman, S. J. *Mol. Ecol.* **6**, 881–885 (1997).

Flame retardants

Persistent pollutants in land-applied sludges

Disposal of sewage sludge by application to agricultural and other land is widely practised and is presumed to be environmentally beneficial, but we have found high concentrations of an environmentally persistent class of organic pollutants, brominated diphenyl ethers (BDEs), in 'biosolids' from four different regions of the United States. These compounds are widely used as flame retardants, and their presence suggests that the environmental consequences of land application of biosolids need further investigation. We also frequently detected BDEs in wild-caught fish, indicating another pathway for human exposure.

Over half of the sewage sludge produced annually in the United States is applied to land, amounting to roughly 4 million tons in 1998 (ref. 1). Sludges are treated before application to reduce odour and pathogen content and their metal burden is regulated. But attention has focused less on persistent organic pollutants since usage of the most notorious (for example, polychlorinated biphenyls) has decreased and pretreatment of industrial waste water has improved^{1,2}.

We analysed 11 biosolid samples before land application from Virginia, Maryland, New York state and California, and found that they all contained high concentrations of BDEs. These flame-retardant polymers are structurally similar to polybrominated biphenyls, the use of which was curtailed after a significant contamination incident in 1973 involving livestock feed in Michigan³.

However, global consumption of BDEs continues to increase, reaching 67,125 metric tonnes in 1999 (refs 4, 5). The most bio-accumulative and toxic BDEs (those containing 4–6 bromine atoms) are being increasingly detected in humans and wildlife from both developed and remote areas^{5–7}. These were present in significant amounts in the biosolids we examined and their relative contributions matched those in 'Penta', the commercial formulation used as a flame retardant in polyurethane foam (Fig. 1). North America accounts for about 98% of global demand for Penta, estimated at 8,290 tonnes in 1999 (ref. 4).

How BDEs are released from polymers has been uncertain, as these applications are considered to be non-dispersive⁷. However, breakdown of discarded polyurethane foam, which may contain up to 30% Penta by weight⁵, may contribute to this. We found that the surface of foam became

brittle and sloughed off after 4 weeks of exposure to ambient summer conditions. The particles generated are easily transported and the polymer matrix preserves the formulation's original BDE composition.

The total concentration of Penta-like BDEs in these biosolids was 1,100–2,290 µg per kg dry weight, suggesting that input was high and consistent, regardless of the region of origin and irrespective of pre-application treatment (see supplementary information). Concentrations exceed those in European sludges by 10- to 100-fold⁸, which is commensurate with the greater demand for Penta in the United States. The European Commission recently proposed a ban on the use of Penta, on the basis of its reported exponential increase in human breast milk and perceived health risks⁹.

The fully brominated Deca product constitutes 82% of the total global BDE market⁴. It is rarely reported in wildlife, perhaps because of its low bioavailability. Deca consists principally of a single BDE (BDE-209) and is used to curtail fires in textiles and in relatively stable, rigid polymers, such as those used in television and computer casings⁵. Unlike those of Penta constituents, BDE-209 concentrations varied widely among the biosolids we analysed (84.8–4,890 µg kg⁻¹; see supplementary information). Although there is little evidence for the degradation of Deca to Penta-like compounds, some photolysis of Deca to less brominated diphenyl ethers is possible^{5,7}.

We also detected BDEs in 87% of fish sampled from Virginia waters (quantification limit in fillets, 5 µg per kg lipid; $n=334$). The principal Penta constituents (BDE-47, -100 and -99) predominated in these samples (Fig. 1). This finding indicates that significant environmental release of these pollutants is occurring in the United States and that humans may be exposed to them through their diet. Carp from one Virginia stream contained 47,900 µg kg⁻¹ of total BDEs, rivalling the highest fillet

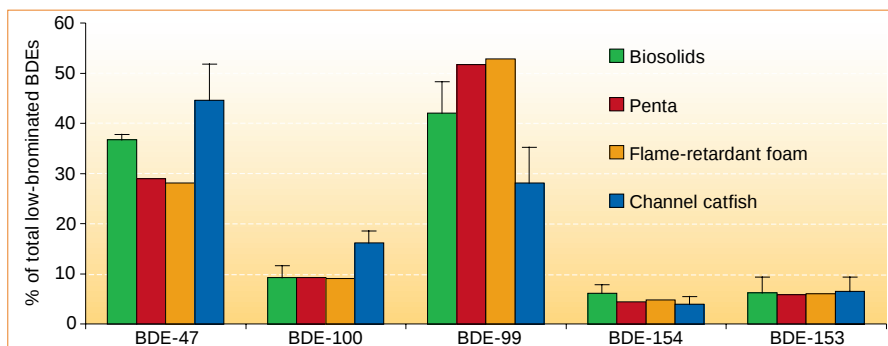


Figure 1 Brominated diphenyl ethers (BDEs) are produced commercially and occur in the environment as mixtures of compounds of varying bromination. Relative contributions of tetra- (BDE-47), penta- (BDE-100 and -99) and hexa- (BDE-154 and -153) brominated versions were similar in 11 biosolids obtained from four different regions of the United States, in the Penta commercial product (used as a flame retardant in polyurethane foam), in treated foam and in wild-caught fish (data shown are for 15 composite samples of channel catfish, *Ictalurus punctatus*, an omnivorous bottom-dwelling species) collected from Virginia lakes and rivers (error bars represent standard deviation). BDE-209 was not detected in fish but was present in biosolids.

burdens reported in the world so far⁴.

These compounds are also detectable in urban and rural air¹⁰, indicating the potential for long-distance atmospheric transport. It seems that BDEs are an important — but generally unrecognized — persistent organic pollutant in the United States. Extensive use of Penta and the high burden of BDEs in land-applied biosolids may facilitate environmental dissemination of less-brominated BDEs both locally and globally.

Robert C. Hale, Mark J. La Guardia, Ellen P. Harvey, Michael O. Gaylor, T. Matteson Mainor, William H. Duff

Department of Environmental Science, Virginia Institute of Marine Science, PO Box 1346, Gloucester Point, Virginia 23062, USA
e-mail: hale@vims.edu

1. US Environmental Protection Agency. *Biosolids Generation, Use and Disposal in the United States*. EPA530-R-99-009 (Washington DC, 1999).
2. National Academy of Sciences. *Use of Reclaimed Water and Sludge in Food Crop Production* (Washington DC, 1996).
3. Di Carlo, F. J. et al. *Environ. Health Perspect.* **23**, 351–365 (1978).
4. Renner, R. *Environ. Sci. Technol.* **34**, 452A–453A (2000).
5. World Health Organization. *Environmental Health Criteria 162: Brominated Diphenyl Ethers* (Geneva, 1994).
6. de Boer, J., Wester, P. G., Klammer, H. J. C., Lewis, W. E. & Boon, J. P. *Nature* **394**, 28–29 (1998).
7. Renner, R. *Environ. Sci. Technol.* **34**, 223A–226A (2000).
8. Sellstrom, U. et al. *Organohalogen Compounds* **40**, 383–386 (1999).
9. European Commission press release, 30 January 2001.
10. Strandberg, B. et al. *Environ. Sci. Technol.* **35**, 1078–1083 (2001).

Supplementary information is available at <http://www.nature.com> or as paper copy from the London editorial office of *Nature*.

Neuroadaptation

Incubation of cocaine craving after withdrawal

Relapse to cocaine addiction is frequently associated with subjective reports of craving, a poorly understood state that precedes and accompanies cocaine-seeking behaviours¹. It has been suggested² that over the first few weeks of withdrawal from cocaine, human addicts become sensitized to drug-associated environmental cues that act as external stimuli for craving, although the evidence for this is inconsistent³. Here we provide behavioural evidence from laboratory animals suggesting that the onset of craving is delayed and that craving does not decay, but rather increases progressively, over a two-month withdrawal period.

We modelled cocaine-craving behaviour by using rats trained to press a lever to receive an intravenous injection of cocaine and then testing them under conditions in which lever-pressing could continue but the cocaine reward was no longer given. In this model, lever-pressing drops to almost zero ('extinguishes') but can be temporarily reinstated by giving the animal an unearned 'priming' injection of the drug⁴, by

administering some forms of stress⁵, or by presenting drug-associated cues⁶ — factors that are known to provoke drug craving in human addicts^{1,7,8}.

We trained seven groups of rats to press the lever for intravenous cocaine injection (0.5 mg per kg body weight per lever-press). Individual rats lived in a chamber that had a retractable lever. Each training session began with insertion of the lever and illumination of a red house light. At the end of each session, the house light was turned off and the lever retracted. A 5-second tone–light signal accompanied each earned injection. After 10 days of 3-hour training sessions twice daily, in which the animals came to earn 55.3 ± 2.7 infusions per day, they were withdrawn from cocaine for 1, 2, 4, 7, 15, 29 or 60 days. During the withdrawal period, the lever was retracted and the house light was kept off.

We subsequently tested each group under two extinction conditions in which cocaine reward was withheld. First, we assessed resistance to extinction in the presence of the house light and the lever — cues that during training had indicated drug availability — but in the absence of the light and tone that were previously paired with drug injection. The animals were allowed to lever-press for six to eight 1-hour sessions (separated by 5-min intervals, during which the lever was retracted and the house light turned out) until their response fell to less than 15 presses per session. We found that lever-pressing was minimal in rats that had been deprived of cocaine for a single day and maximal in animals that had been deprived for 60 days (Fig. 1a).

The second test of cocaine seeking was a cue-induced reinstatement test conducted 5 min after the last of the extinction sessions. This test began with a 5-second presentation of the tone–light signal that had previously accompanied cocaine injection; each lever-press in this test resulted in another presentation of the tone–light signal⁶. In this test, the animals were not only exposed to the cues that would normally signal cocaine availability, but they were also exposed to the conditioned reinforcing cues that previously confirmed cocaine reward. Again, response was minimal after a single day of cocaine deprivation and maximal after 60 days (Fig. 1b). We found a linear increase over 2 months of cocaine withdrawal in the rats' sensitivity to similar drug-associated environmental cues that stimulate cocaine craving in humans¹.

Our results are consistent with clinical observations in humans² and suggest that a delayed-onset craving syndrome develops or 'incubates' during the first 2 months of cocaine abstinence, and probably lasts for longer. Although the mechanisms responsible for this incubation are not known, the intensification of cocaine seeking described

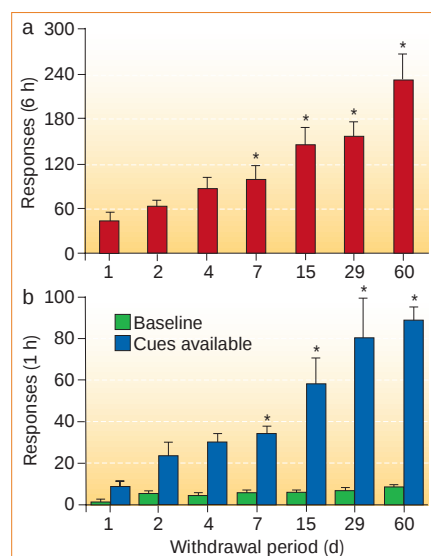


Figure 1 Persistence of a cocaine-seeking habit as a function of time since the last day of self-administration of cocaine. **a**, Mean ($\pm$ standard error) number of non-reinforced responses on the lever previously associated with cocaine, from six extinction sessions in the presence of the house light and lever cues that were previously associated with cocaine availability. **b**, Mean ($\pm$ standard error) number of non-reinforced responses on the lever previously associated with cocaine in the subsequent presence of the light–tone signal (conditioned reinforcer) that was previously associated with earned cocaine injections. Baseline data are from the previous extinction session. *Different from day 1 ($P < 0.01$).

here develops over a period when most of the neuroadaptations that accompany withdrawal from chronic cocaine addiction are in progressive decline^{9–11}.

The time course of this intensified drug seeking is similar to that of psychostimulant sensitization, which becomes progressively stronger with increasing abstinence for periods of up to several weeks^{12,13}. Whatever the mechanism by which craving is incubated, our evidence is inconsistent with the view that cocaine craving decays progressively after cessation of drug use. It suggests instead that the individual is most vulnerable to relapse at times well beyond the acute phase of drug withdrawal.

Jeffrey W. Grimm, Bruce T. Hope, Roy A. Wise, Yavin Shaham

Behavioral Neuroscience Branch, Intramural Research Program, National Institute on Drug Abuse, National Institutes of Health, 5500 Nathan Shock Drive, Baltimore, Maryland 21224, USA
e-mail: yshaham@intra.nida.nih.gov

1. Childress, A. R. et al. *Am. J. Psychiatry* **156**, 11–18 (1999).
2. Gawin, F. H. & Kleber, H. D. *Arch. Gen. Psychiatry* **43**, 107–113 (1986).
3. Satel, S. L. et al. *Am. J. Psychiatry* **148**, 1712–1716 (1991).
4. Stewart, J. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **7**, 591–597 (1983).
5. Shaham, Y., Erb, S. & Stewart, J. *Brain Res. Rev.* **33**, 13–33 (2000).
6. Grimm, J. W. & See, R. E. *Neuropsychopharmacology* **22**, 473–479 (2000).
7. de Wit, H. *Exp. Clin. Psychopharmacol.* **4**, 5–10 (1996).
8. Sinha, R., Fuse, T., Aubin, L. R. & O'Malley, S. S. *Psychopharmacology* **152**, 140–148 (2000).
9. Nestler, E. J. *Nature Rev. Neurosci.* **2**, 119–128. (2001).

10. White, F. J. & Kalivas, P. W. *Drug Alcohol Depend.* **51**, 141–153 (1998).
11. Sanyal, Z., Shahan, Y. & Heinrichs, S. C. *Pharmacol. Rev.* (in the press).
12. Vanderschuren, L. J. & Kalivas, P. W. *Psychopharmacology* **151**, 99–120 (2000).
13. Robinson, T. E. & Berridge, K. C. *Brain Res. Rev.* **18**, 247–291 (1993).

Neuropharmacology

Odorants may arouse instinctive behaviours

The prevailing view of the mammalian olfactory system is that odorants are detected only in the nasal olfactory epithelium, whereas pheromones are generally detected in the vomeronasal organ^{1–3}. Here we show that vomeronasal neurons can actually detect both odorants and pheromones. This suggests that in mammals, as in insects^{4–6}, odorous compounds released from plants or other animal species may act as ‘semiochemicals’ — signalling molecules that elicit stereotyped behaviours that are advantageous to the emitter or to the receiver.

To investigate the function of the vomeronasal organ, we used calcium imaging of single murine vomeronasal neurons containing Fura-2 dye^{7,8}. As with olfactory neurons in the nose, resting concentrations of intracellular calcium were about 20–40 nM in vomeronasal neurons and were increased to about 120–150 nM by 100 mM potassium chloride⁷. We found that vomeronasal neurons from both males and females respond to six mouse pheromones that stimulate aggression, subordination or alteration in puberty onset or oestrus (Fig. 1a, b)⁹. Individual pheromones (100 µM) stimulated 0.3–0.7% of the neurons, most of which responded to only one pheromone, as shown previously¹⁰.

Surprisingly, mouse vomeronasal neurons also detect odorants (Fig. 1; Table 1). We assembled 82 odorants (50 µM each) in 9 mixes and found that neurons responded to several of these mixes. Vomeronasal neurons that responded to an individual mix were tested against each odorant in the mix. Altogether, 0.1–1.5% of neurons responded to a single mix (Table 1). Neurons were also

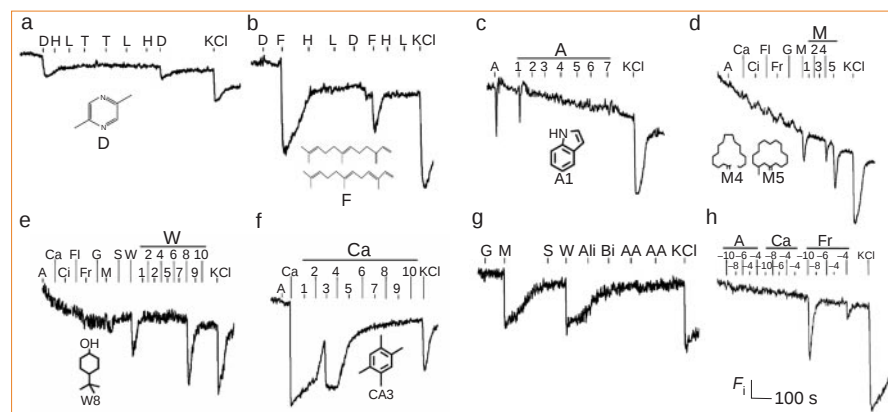


Figure 1 Responses of single vomeronasal neurons (VNs) to pheromones and odorants. Calcium imaging⁷ was used on dissociated VNs containing Fura-2 during exposure by perfusion to pheromones (a, b) or odorant mixes/odorants (c–h) (4 seconds each) and then to 100 mM KCl. Fluorescence emission (at 510 nm) from cells illuminated at 380 nm was monitored (F_i, fluorescence intensity in arbitrary units)⁷. Scale is the same for all traces. The lack of response to all concentrations (10^{−10}–10^{−4} M) of an odorant in h may be due to transient desensitization. Odorant (and other) mixes: animalic (A), camphoric (Ca), citrus (Ci), floral (Fl), fruity (Fr), green minty (G), musky (M), sweet (S), woody (W), aliphatic (Al) and amino acids (AA). Pheromones: dehydro-exo-brevicomin (B), 2-heptanone (H), 2,5-dimethyl pyrazine (D), 2-sec-butyl-4,5-dihydrothiazole (T), E,E-α-farnesene plus E-β-farnesene (F) and lactol (L). Single odorants that elicited responses: indole (A1), hexadecanolide (M4), muscone (M5), patchouli (W8) and durenene (Ca3); as well as (not shown) p-cresol (A2), eucalyptol (Ca1), isoborneol (Ca4), borneol (Ca5), fenchone (Ca6), butyrolphenone (Ca7), methylanisole (Ca8), myrtenal (Ca9), phenylacetaldehyde (F15), dimethyl-3-octanol (F18), helional (F110), pentadecalactone (M3) and aubepine (S1).

activated by 18 single odorants classified as animalic, camphoric, floral, musky, sweet or woody (Fig. 1).

We found that, like olfactory neurons^{7,8}, vomeronasal neurons were activated by more than one odorant or mix, but they can also distinguish between highly related odorants such as indole and skatole, which differ by a single methyl group (Fig. 1c). Like pheromones¹⁰, the three odorant mixes tested activated vomeronasal neurons (n = 6) at 10^{−10} M (Fig. 1h), which is much less than is required for an olfactory neuron to respond, indicating that the vomeronasal organ is highly sensitive to low concentrations of both pheromones and odorants.

Vomeronasal neurons detected many of the odorants we tested (18 out of 82). It is unlikely that so many odorants would be released from mice as pheromones. But why does the vomeronasal organ detect odorants? In contrast to the olfactory epithelium, there is no direct pathway from this organ to the higher cortical areas involved in odour perception and discrimination^{1–3}. Instead, inputs are targeted to the amygdala and hypothalamus, areas that control hormone levels, emotions, basic drives and

instinctive behaviours. Like pheromones, some odorants may stimulate innate behavioural or physiological responses.

As in insects, certain odorants may act in mammals as semiochemicals that influence behaviour. Volatile chemicals emitted by plants can elicit oviposition or pollination in insects, and those released from prey can stimulate prey-finding behaviours^{4–6}. A prey protein detected by the vomeronasal organ of the garter snake also induces tracking activity¹. Certain odorants in the natural habitat of mice may similarly provide cues that signal the presence of a predator or indicate the suitability of a particular site for feeding or nesting.

Mehran Sam*, Sadhna Vora*, Bettina Malnic*, Weidong Ma†, Milos V. Novotny†, Linda B. Buck*

*Howard Hughes Medical Institute, Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA

e-mail: lbuck@hms.harvard.edu

†Institute for Pheromone Research, Department of Chemistry, Indiana University, Bloomington, Indiana 47405, USA

1. Halpern, M. *Annu. Rev. Neurosci.* **10**, 325–362 (1987).
2. Wysocki, C. J. & Meredith, M. in *The Neurobiology of Taste and Smell* (eds Finger, T. E. & Silver, W. L.) 125–150 (Wiley, New York, 1987).
3. Buck, L. B. *Cell* **100**, 611–618 (2000).
4. Binder, B., Robbins, J. & Wilson, R. J. *Chem. Ecol.* **21**, 1315–1327 (1995).
5. Dobson, H., Groth, I. & Bergstrom, G. *Am. J. Bot.* **83**, 877–885 (1996).
6. Kielty, J., Allen-Williams, L., Underwood, N. & Eastwood, E. *J. Insect Behav.* **9**, 237–250 (1996).
7. Sato, T., Hirono, J., Tonoike, M. & Takebayashi, M. *J. Neurophysiol.* **72**, 2980–2989 (1994).
8. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. *Cell* **96**, 713–723 (1999).
9. Novotny, M., Jemiolo, B. & Harvey, S. in *Chemical Signals in Vertebrates* (eds Macdonald, D. W., Muller-Schwarze, D. & Natynczuk, S. E.) 1–21 (Oxford Univ. Press, Oxford, 1990).
10. Leinders-Zufall, T. et al. *Nature* **405**, 792–796 (2000).

Table 1 Responsiveness of vomeronasal neurons to odorants

Odorant mix	Neurons tested*	Responsive neurons (%)
Animalic	1,045	13 (1.2)
Camphoric	973	15 (1.5)
Citrus	731	2 (0.3)
Floral	719	4 (0.6)
Fruity	848	5 (0.6)
Green minty	696	1 (0.1)
Musky	696	4 (0.6)
Sweet	626	3 (0.5)
Woody	596	4 (0.7)

*Number of KCl-responsive vomeronasal neurons tested with each odorant mix. Some neurons were tested with all mixes and others with a subset of mixes.

Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats

M. F. Perutz* & A. H. Windle†

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

†University of Cambridge, Department of Materials Science and Metallurgy, Pembroke Street, Cambridge CB2 3QZ, UK

Neurodegenerative diseases resulting from expanded repeat sequences of glutamine residues are associated with the formation of protein aggregates in the cell nuclei of the affected neurons, but whether these are pathogenic is controversial. Recent observations indicate that the ages of onset of these diseases are exponential functions of the repeat lengths and that the probability of neural death is constant with time. The only process known to us that could give rise to such behaviour is nucleation of the aggregates.

Huntington's chorea is a dominantly inherited, late-onset, neurodegenerative disease that is associated with the expansion of a glutamine repeat near the amino terminus of a protein more than 3,140 amino-acid residues long. The greater the expansion, the earlier the start of the disease. The normal function of the protein is unknown and its amino-acid sequence shows no homology with any known protein. The length of the glutamine repeat varies: in healthy individuals it is less than 37, but individuals with more than 40 repeats never remain free from the disease. The same is true of six out of seven other neurodegenerative diseases¹, each of which is due to expansion of a glutamine repeat in a different protein.

Synthetic poly-L-glutamine aggregates into pleated β -sheets that are linked by hydrogen bonds between both their main-chain and side-chain amides². This adhesive property causes the proteins that possess expanded glutamine repeats to form aggregates in the nuclei of affected neurons^{3,4}. However, it has not yet been established whether these aggregates are themselves the cause of neural death.

Two recent observations make an important contribution to this discussion: one is the constant probability of neural death with time in Huntington's disease⁵; the other is the finding that in all eight neurodegenerative diseases caused by expansion of glutamine repeats, the age of onset is an exponential function that is inversely dependent on the length of the repeats¹. Theoretical considerations show that these findings are consistent with the aggregates being the cause of cell death.

Neural death over time

Delays in the clinical onset of Huntington's chorea and other late-onset neurodegenerative diseases are often attributed to cumulative cell damage; Clarke *et al.*⁵ point out that this would lead to an increased probability of neural death with time, but their statistical analyses of neural deaths in 12 such diseases show no such increase. In Huntington's disease, the probability of neural death remains constant with time. Clarke *et al.* account for this constancy by proposing a single-hit model in which the death of a neuron is initiated randomly in time by a single, rare catastrophic event⁵. The best-known example of such a random process is radioactive decay. In solutions of macromolecules, such as those that exist in living cells,

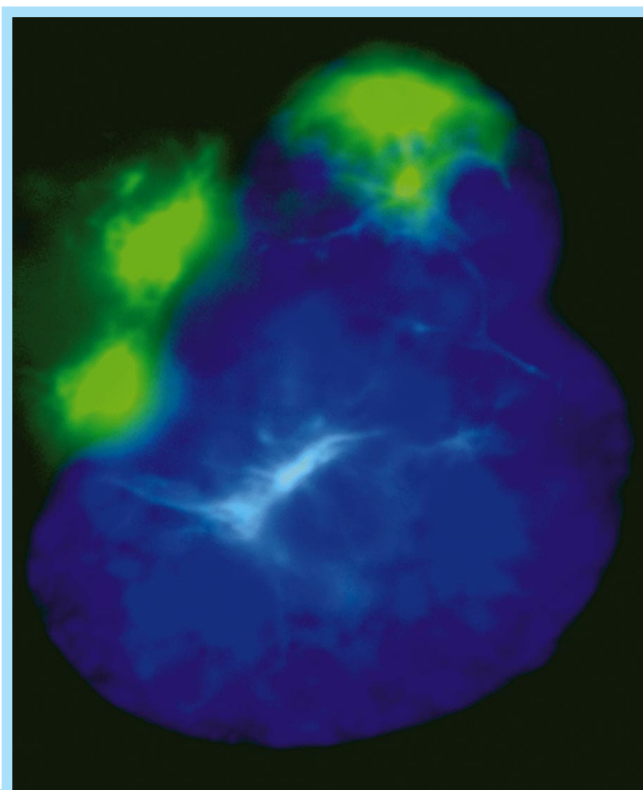


Figure 1 Aggregation from several nuclei of the exon-1 product of the Huntington's disease gene with a repeat of 47 glutamine residues, tagged with a fluorescent protein of relative molecular mass 26,000, and expressed in a single COS-1 cell. Note the fibrous network. Image courtesy of A. Kazantsev, E. Preisinger and D. Ho.

the only random mechanism of this kind known to us is nucleation of ordered molecular aggregates.

Physics of aggregation

Nuclei of molecular aggregates are unstable when very small because of their large ratio of surface area to volume, coupled with the energy

penalty associated with the surface. Their surface energy thus overrides the greater stability of the aggregate (per unit volume) compared with that of the surrounding solution. The principle is enshrined in the Gibbs–Thomson equation:

$$kT \ln(p_r/p_0) = 2\sigma v/r$$

where p_r/p_0 is the ratio of vapour pressure over a spherical droplet of radius r to that over a flat surface, σ is the surface energy, and v is the molecular volume (ref. 6). For precipitation from solution, p_r/p_0 can be replaced by S , the supersaturation ratio. For a given value of S , there will be a critical radius, r_c , that is just sufficient to achieve stability, at which the probability of further growth is greater than the probability of shrinkage towards complete redissolution.

The critical free energy required to create a spherical nucleus of critical radius, then, is

$$\Delta G_{\text{crit}} = \frac{4\pi\sigma r_c^2}{3} = \frac{16\pi\sigma^3 v^2}{3(kT \ln S)^2}$$

where σ is the surface energy of the precipitate, v is the molecular volume, S is the supersaturation ratio, and kT (Boltzmann's constant $\times$ absolute temperature) describes the mean thermal energy of the system. The only way in which this free energy could be supplied is by a chance yet collaborative organization of a sufficient number of molecular units. The probability of this free energy being reached, bearing in mind that kT is the mean of a random distribution of thermal energy, is proportional to $\exp[-\Delta G_{\text{crit}}/kT]$.

The addition of an appropriate pre-exponential factor, A , provides an expression of the rate of formation of stable nuclei (A is typically $10^{31} \text{ m}^{-3} \text{ s}^{-1}$). The formation of nuclei is therefore a random process in time and space. Because of the exponential, the rate is heavily dependent on the physical parameters of surface energy, molecular volume, temperature and (somewhat less heavily) supersaturation of the surrounding solution.

The actual mechanism of the generation of nuclei based on polyglutamine sequences will be structurally complex, but the balance between volume, free energy and surface energy, and the way in which this balance determines the size and hence the probability of creation of a critical nucleus for further sustained growth is an underlying principle. Aggregates should appear at random intervals of time, and the time constant will depend, among other factors, on the length of the repeats in the protein. In Huntington's chorea, which takes decades to develop, it is clearly a rare event. On the other hand, the growth rate of a precipitate, once nucleated, in a range of supersaturations that would allow nucleation on a non-geological timescale, will depend on the availability of monomeric Huntington's protein, but is typically of the order of tens of micrometres per second. For any process that occurs on a timescale of years, the controlling step will be nucleation, not growth, and it will occur at random intervals of time.

Before any one Huntington's protein molecule aggregates, transition to a new intramolecular configuration must occur. Such transitions may take place even in isolated molecules in solution, but they, or altered interactions with other molecules, cannot be the rate-determining steps in neural death, because they take only fractions of a second, whereas nucleation, like neural death, may take tens of years. If the production of the Huntington's protein were terminated and some scavenging action, say by proteasomes, degraded all these proteins in solution in the cell, then thermodynamics suggests that any existing aggregates would tend to redissolve. Like radioactive elements, neurons would have a characteristic half-life: in Huntington's chorea, their average half-life is reduced by a factor of 1.055 for every glutamine residue added to the repeat.

Conclusions

In cell or animal models, investigators have reported cell death without aggregates and aggregates without cell death^{7–9}, but these results are open to alternative interpretation^{10,11}. They are also called into question by the discovery that aggregates kill neurons by binding to, and slowly depleting them of, an essential transcriptional co-activator that contains a repeat of 18 glutamine residues, the CREB-binding protein¹².

Several other observations support our theoretical expectations. In COS cells transfected with the exon-1 product of the Huntington's gene, nucleation of the protein occurs randomly and is followed by rapid aggregation. The frequency of nucleation and the rate of growth of the aggregates increase with the number of glutamine repeats and the amount of expression of the protein^{13,14}. In transgenic mice made homozygous for the entire Huntington's gene with extended CAG repeats, symptoms appear earlier than in heterozygotes¹³. In humans, the difference between homo- and heterozygotes cannot be judged, as it falls within the range of variation in onset age associated with heterozygotes¹³. In transgenic mice, the formation of nuclear inclusions precedes the development of symptoms, whereas inhibition of the expression of the Huntington's gene causes regression of both aggregates and symptoms¹⁵.

The number of repeats that causes aggregation of the Huntington's protein *in vitro* is the same as the number that causes the disease^{16,17}. Most neurodegenerative conditions, such as Alzheimer's and Parkinson's diseases, prion and other amyloid diseases, are associated with protein aggregates in neurons. There is disagreement over whether extracellular senile plaques or intracellular double-helical filaments are responsible for Alzheimer's disease, but the formation of aggregates is undoubtedly linked with these diseases, although we do not always understand why they are toxic. It would be surprising if Huntington's chorea were an exception. We suggest, therefore, that in all these diseases, therapy should be aimed at preventing or reversing aggregation.

Received 6 November 2000; Accepted 14 June 2001.

- Gusella, J. F. & MacDonald, M. E. Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. *Nature Rev. Neurosci.* **1**, 109–115 (2000).
- Perutz, M. F., Johnston, T., Suzuki, M. & Finch, J. T. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl Acad. Sci. USA* **91**, 5355–5358 (1994).
- Davies, S. W. *et al.* Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell* **90**, 537–548 (1997).
- DiFiglia, M. *et al.* Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science* **277**, 1990–1993 (1993).
- Clarke, G. *et al.* A one-hit model of cell death in inherited neuronal degenerations. *Nature* **406**, 195–199 (2000).
- Mullins, J. W. *Crystallisation* 2nd edn (Butterworth, London, 1972).
- Skinner, P. J. *et al.* Ataxin-1 with an expanded glutamine tract alters nuclear matrix-associated structures. *Nature* **389**, 971–974 (1997).
- Saudou, F., Finkbeiner, S., Devys, D. & Greenberg, M. E. Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear aggregates. *Cell* **95**, 55–66 (1998).
- Kuemmerle, S. *et al.* Huntingtin aggregates may not predict neuronal death in Huntington's disease. *Neurology* **46**, 842–849 (1999).
- Carmichael, J. *et al.* Bacterial and yeast chaperones reduce both aggregate formation and cell death in mammalian cell models of Huntington's disease. *Proc. Natl Acad. Sci. USA* **97**, 9701–9705 (2000).
- Perutz, M. F. Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends Biochem. Sci.* **24**, 58–63 (1999).
- Nucifora, F. C. Jr *et al.* Interference by huntingtin and atrophin-1 with CPB-mediated transcription leading to cellular toxicity. *Science* **291**, 2423–2428 (2001).
- Narain, Y., Wittenbach, A., Rankin, J., Furlong, R. A. & Rubinsztein, D. C. A molecular investigation of true dominance in Huntington's disease. *J. Med. Genet.* **36**, 739–746 (1999).
- Kazantsev, A. *et al.* A polypeptide inhibitor of aggregation suppresses pathogenesis in a *Drosophila* model of polyglutamine repeat disease. *Cell* (submitted).
- Yamamoto, A., Lucas, J. J. & Hen, R. Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell* **101**, 57–66 (2000).
- Scherzinger, E. *et al.* Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates *in vivo* and *in vitro*. *Cell* **90**, 549–558 (1997).
- Huang, C. C. *et al.* Amyloid formation by mutant huntingtin: threshold, progressivity and recruitment of normal polyglutamine proteins. *Somat. Cell Mol. Genet.* **24**, 217–233 (1998).

An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30° N

Deborah S. Kelley*, Jeffrey A. Karson†, Donna K. Blackman‡, Gretchen L. Früh-Green§, David A. Butterfield*, Marvin D. Lilley*, Eric J. Olson*, Matthew O. Schrenk*, Kevin K. Roell, Geoff T. Lebon||, Pete Rivizzigno† & the AT3-60 Shipboard Party

* University of Washington, School of Oceanography, Seattle, Washington 98195, USA

† Duke University, Division of Earth & Ocean Sciences, Durham, North Carolina 27708-0230, USA

‡ Scripps Institution of Oceanography, La Jolla, California 92093-0225, USA

§ Institute For Mineralogy and Petrology, ETH-Zentrum, CH-8092, Zurich, Switzerland

|| Joint Institute for the Study of the Atmosphere & Ocean, University of Washington and NOAA Pacific Marine Environmental Laboratory, Seattle, Washington 98115, USA

Evidence is growing that hydrothermal venting occurs not only along mid-ocean ridges but also on old regions of the oceanic crust away from spreading centres. Here we report the discovery of an extensive hydrothermal field at 30° N near the eastern intersection of the Mid-Atlantic Ridge and the Atlantis fracture zone. The vent field—named ‘Lost City’—is distinctly different from all other known sea-floor hydrothermal fields in that it is located on 1.5-Myr-old crust, nearly 15 km from the spreading axis, and may be driven by the heat of exothermic serpentinization reactions between sea water and mantle rocks. It is located on a dome-like massif and is dominated by steep-sided carbonate chimneys, rather than the sulphide structures typical of ‘black smoker’ hydrothermal fields. We found that vent fluids are relatively cool (40–75 °C) and alkaline (pH 9.0–9.8), supporting dense microbial communities that include anaerobic thermophiles. Because the geological characteristics of the Atlantis massif are similar to numerous areas of old crust along the Mid-Atlantic, Indian and Arctic ridges, these results indicate that a much larger portion of the oceanic crust may support hydrothermal activity and microbial life than previously thought.

Most known hydrothermal fields along mid-ocean ridges are located on young crust where the cooling of hot basaltic material drives hydrothermal flow¹. In such systems, precipitation of iron- and sulphide-rich minerals occurs during mixing of 200–400 °C hydrothermal fluids with cold, oxygenated sea water. The compositions of the resulting sulphide chimneys reflect fluid–rock reactions within the underlying basaltic–gabbroic substrate^{2,3}. All evidence indicates that such black smoker systems and associated diffuse flow typify hydrothermal activity directly on-axis in mid-ocean-ridge environments. However, there is a growing body of evidence from recent water column and sea-floor studies indicating that lower-temperature venting associated with older, tectonized portions of the oceanic crust may be common along much of the mid-ocean-ridge spreading network^{4–6}. Here we describe the Lost City hydrothermal field, which represents the first observation of this type of low-temperature venting associated with extensive chimney development. The Lost City field is spectacular in that it hosts numerous actively venting structures, one of which reaches 60 m in height. The steep-sided pinnacles are composed entirely of carbonate and magnesium hydroxide minerals, making them distinctly different from other well known mid-ocean-ridge hydrothermal vents.

Geology and tectonic setting of the Lost City hydrothermal field

The Atlantis massif is located at the inside corner of the intersection of the Mid-Atlantic Ridge (MAR) and the 75-km, left-lateral offset, Atlantic transform fault (ATF) (Fig. 1)^{7,8}. The massif is approximately 15 km across and the southern flanks are steep escarpments with 3,800 m of relief adjacent to the ATF. The upper surface of the dome is interpreted as a major low-angle normal or detachment fault^{7,8} that has exposed variably metamorphosed peridotite and gabbro. The top of the scarp is marked by a sharply defined unconformity overlain by a laterally variable assemblage of very gently dipping, undeformed sedimentary rocks. These include carbonate cemented breccias with clasts of basalt, gabbro and peridotite, and well lithified, bedded carbonates. These are overlain by variably consolidated pelagic ooze with dispersed blocks and

rubble of basaltic and ultramafic material. The unconformity indicates that the southern end of the massif may have been near or even above sea-level before subsiding to its current depth of about 700 m. The south wall of the massif is a series of steep cliffs that define an extensive, south-facing embayment with several steep-sided ridges that extend southward toward the transform valley (Fig. 1). Magnetic anomaly patterns show that the centre of the massif, about 15 km west of the spreading axis, is about 1.5 Myr old, consistent with the local half-spreading rate of 12 mm yr^{−1} (ref. 8).

The Lost City field

Investigations of the southern wall of the massif, using the remotely operated imaging vehicle *ArgoII* and the submersible *Alvin*, resulted in the discovery of the Lost City field (LCF). The LCF rests on a terrace at a water depth of 700–800 m on a south-trending spur that protrudes from the crest of the south wall scarp (Fig. 1). The field is underlain by a diverse suite of mafic and ultramafic rocks that crop out on the cliffs immediately below the edge of the scarp. These include a complex assemblage of variably serpentinized and deformed peridotites, massive gabbro to oxide gabbro, and meta-gabbros. The hydrothermal structures and related deposits overlie and fill fractures in the capping carbonate unit and thus clearly post-date this assemblage. In addition, the vent structures lack pelagic sedimentary cover, suggesting that they may be relatively young.

Initial surveys of the surrounding area with *ArgoII* and *Alvin* indicate that the field extends for at least 400 m across the terrace and that it hosts at least 30 active and inactive structures (Fig. 2). To the south, cliff exposures have extensive areas of active and inactive white hydrothermal precipitates. These deposits fill fractures and form hundreds of shelf-like overhanging ledges or ‘flanges’ that protrude as much as 2 m from the cliff face.

Within the LCF, active and inactive vents exhibit a wide variety of morphologies that include small spires, mounds and pinnacles (Fig. 2a). The mounds are variably cemented, 10–20 m high, steep-sided deposits composed of small toppled spires that have been overgrown and cemented by hydrothermal precipitates. Larger

isolated pinnacles are commonly 10–30 m tall. But the most spectacular of the pinnacles is a giant columnar tower that rises 60 m above the sea floor, making it the tallest hydrothermal deposit yet discovered anywhere on the sea floor. The top of this composite structure is 15 m across and hosts four smaller spires. At least one of these spires is actively venting 75 °C fluid from its top. On this and other large pinnacles, flanges exhibit concave-down forms, which trap highly reflective pools of 40–55 °C vent fluid (Fig. 2b). Delicate fingers and dendritic growth ornament the edges and tops of the flanges; stalagmite-like cones rise several metres from the tops of some flanges. In most cases, fresh-looking white deposits indicate

that active venting occurs on the tops of the spires and pinnacles as well as from the flanges.

Mineral and fluid chemistry

X-ray diffraction analyses of samples from seven inactive and active chimneys and flanges indicate that the structures are composed of variable mixtures of calcite (CaCO_3), aragonite (CaCO_3), and brucite ($\text{Mg}(\text{OH})_2$). The composition of the structures is reflected in the chemistry of three vent fluid samples collected with *Alvin* from a 40 °C and 75 °C site. The pH of the vent fluids measured at 25 °C is high (9.0 to 9.8 versus 8.0 for ambient sea water) (Table 1).

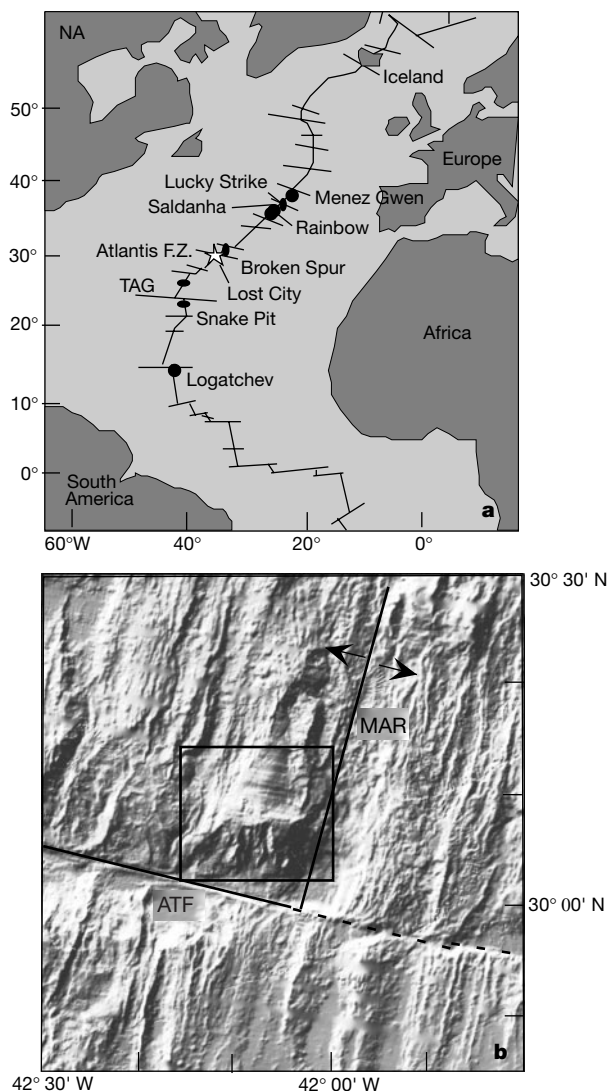


Figure 1 The Mid-Atlantic Ridge and location of the Lost City field. **a**, Location of active hydrothermal sites along the Mid-Atlantic Ridge (dots) and the Lost City hydrothermal field on the Atlantis massif at 30° N. In addition to the Lost City, the Logatchev, Rainbow and Saldanha fields are also hosted on peridotite and gabbroic material. Saldanha most closely resembles the LCF in that it is also located on a peridotite massif at a water depth of 700 m, it hosts filamentous bacteria, and no vent fauna were identified. Venting of clear, warm fluids was observed from small orifices through sediment⁶. **b**, Shaded relief map showing the location of the Atlantic massif and the study site (box). Also shown is the location of the Mid-Atlantic Ridge (MAR) and the Atlantis transform fault (ATF). The southern face of the massif is a steep sided scarp with nearly 3,800 m of relief. The hydrothermal field is located at a water depth of about 700 m near the top of the massif. The dotted line denotes the trace of the ATF.

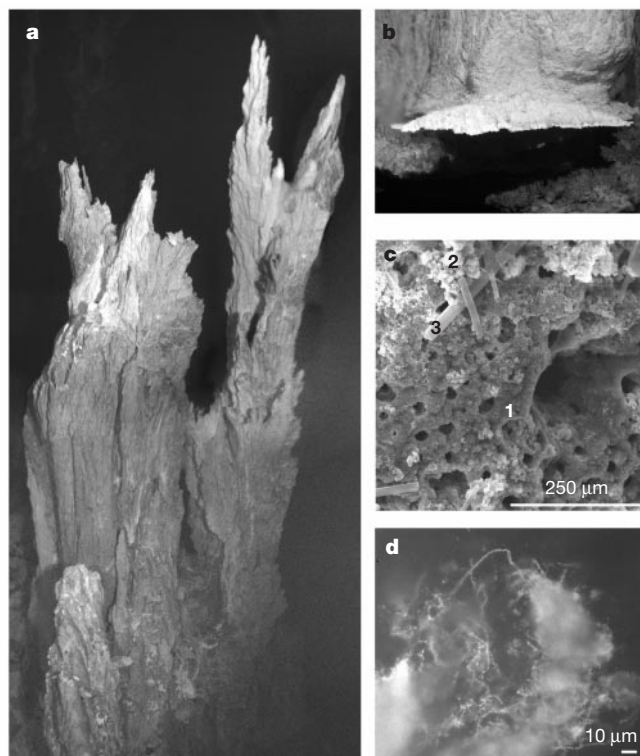


Figure 2 Hydrothermal deposits and microbial communities within the Lost City field. **a**, Photomosaic of an inactive 8-m-tall carbonate chimney in the eastern portion of the Lost City field. This mosaic was produced from digital still camera imagery collected every 15 s by the remotely operated vehicle *ArgoII*. The calcite, aragonite and brucite chimneys form delicate to massive pinnacles that reach up to 60 m in height. **b**, Aragonite and brucite flange venting 40 °C fluids (shimmering water in left portion of the image). The carbonate ledges grow horizontally out from the chimney walls and trap buoyant reflecting pools of warm water, which seeps out from the lip of the flanges, and up through porous flange tops results in outward growth and thickening of the flanges. The flange shown in this image is about 1 m in width and hosts abundant microbial communities. **c**, Scanning electron image (SEM) of a piece of the flange shown in **b**, collected with *Alvin*. Elemental detection and X-ray diffraction analyses of this sample show that it contains a fine porous matrix of calcium carbonate (aragonite) (point 1), and magnesium hydroxide ($\text{Mg}(\text{OH})_2$) minerals (points 2 and 3), which exhibit variable morphologies. The SEM used was an ISI DS-130s with an operating voltage of 18 kV. The images were collected using ITRF Iridium II EDS software. Molecular ratios were determined using ZAF (atomic number, absorption, fluorescence) corrections after deconvolution through the ITRF software. Specimens were sputter coated with palladium before being analysed. **d**, Epifluorescent microphotograph of DAPI (4',6-diamidino-2-phenylindole)-stained filamentous microbial communities in the flange sample collected from the site shown in **b**. Continuous biofilms composed of several types of microbial cells were observed attached to mineral surfaces within the active vent structures. Microbial cells ranged from 0.5 to 2.0 μm in diameter and included cocci, rods and filaments. Significant biomass is observed within the active samples recovered.

These values are in marked contrast to vent fluids collected from basalt-hosted environments where pH values measured at 25 °C are typically 3.0 to 5.0 (Table 1). High pH and Ca concentration are typical of fluids emanating from serpentinized ultramafic rocks^{9,10} and promote carbonate precipitation upon mixing with sea water.

Lost City fluids have magnesium concentrations much lower than sea water (9–19 mmol kg⁻¹ versus 54 mmol kg⁻¹) and the three samples show nearly linear mixing trends when other major elements are plotted against Mg. As there is brucite in the chimneys and flanges, Mg is probably reactive within the edifices. Mg²⁺ activity derived from sea water in the fluids may favour the precipitation of aragonite, which is present in many of the Lost City samples and is commonly found associated with serpentinized peridotites^{11–15}. Ca is enriched more than twofold in the fluids, but K is within 3% of the ambient seawater value. The 75 °C fluids have slightly higher Ca concentration at zero Mg than the cooler samples, consistent with progressive seawater mixing and carbonate precipitation. Reactive silicate (measured after two weeks of refrigerated storage) is lower than ambient sea water in the 40 °C samples, but is slightly higher than sea water in the 75 °C sample. Silica is a trace component within the chimney minerals. Over 60 μmol kg⁻¹ of total H₂S was detected after nine days of sample storage, and SO₄ is in excess over values predicted from mixing sea water and an upwelling end-member with zero Mg and SO₄. Na and Cl are both within 1% of the sea-water value (Table 1). Measured hydrogen and methane concentrations were 249–428 and 136–285 mmol kg⁻¹, respectively.

The LCF is the first known example of sea-floor vents capable of producing the low Mn/CH₄ plumes that are common along fracture zones and non-transform offsets of the MAR. The high Ca, low Mg and near-ambient silica content are consistent with peridotite-dominated fluid–rock interaction^{16,17} producing an alkaline fluid that precipitates carbonates and hydroxides below the sea floor and upon mixing with sea water.

Interaction of mantle materials with sea water during serpentinization is further supported by stable isotope analyses of carbonate in the pinnacle structures. For the seven samples recovered, the δ¹³C values range from 1.0 to 2.1‰ (VPDB, the Vienna Pee Dee belemnite standard) and clearly reflect a marine source of

carbon¹⁴. The δ¹⁸O values of 32.5 to 35.4‰ (VSMOW, Vienna Standard Mean Ocean Water) are slightly enriched in ¹⁸O relative to marine carbonates, but they are typical values for aragonite associated with oceanic serpentinites (31–36‰ VSMOW)^{12,14,15}. Calculations based on published oxygen isotope fractionation factors and a temperature of 7 °C (as measured for the ambient bottom water in the area) indicate that most of the pinnacle and flange carbonates were precipitated from altered sea water with δ¹⁸O values of 0.5 to 2‰ (VSMOW). These values are consistent with δ¹⁸O values of serpentinizing fluids calculated from serpentine oxygen isotope data from different tectonic environments^{13–15} and suggest that active subsurface serpentinization reactions below the massif control the oxygen isotope signatures of the fluids venting in the LCF.

Life within the vent system

Within this field, the active carbonate chimneys are typically awash in buoyantly rising mixtures of warm shimmering vent fluid and cooler sea water. These diffusely venting areas support dense microbial communities that commonly form white to light grey coloured filamentous strands several centimetres in length. Preliminary investigations of microbial communities show extensive biofilm development on mineral surfaces within the carbonate structures (Fig. 2d). Samples obtained from a 75 °C site at the top of the 60-m-tall structure contain abundant biomass, which exists primarily as microcolonies and isolated cells on the surfaces of carbonate minerals (Fig. 2d). Enrichment culturing of chimney material in aerobic and anaerobic media yielded microorganisms in the thermophilic (50 °C, 70 °C) and mesophilic (25 °C) temperature regimes. Preliminary results of DNA extraction and analysis from recovered flange and chimney material indicates that Archaeal and Eubacterial lineages are both present at Lost City. However, macrofaunal assemblages that typify most vent environments¹⁸ are extremely rare within the LCF and are limited to a few crabs, sea urchins, and abundant sponges and corals.

Global significance of off-axis venting

During the past two decades hydrothermal fields have been explored at over 40 sites along the mid-ocean-ridge spreading network^{1,6,19–21}. Eight of these occur along the axis of the MAR and three are hosted

Table 1 Summary of vent fluid data

Location	Host rock	T (°C)	pH	Mg (mmol kg ⁻¹)	Ca (mmol kg ⁻¹)	Na (mmol kg ⁻¹)	Cl (mmol kg ⁻¹)	SO ₄ (mmol kg ⁻¹)	H ₂ S (mmol kg ⁻¹)	CH ₄ (mmol kg ⁻¹)	H ₂ (mmol kg ⁻¹)	Reference
Sea water		7	8.0	54.0	10.4	475	553	28.6	0	4 × 10 ⁻⁷	4 × 10 ⁻⁴	
Lost City; 30° N MAR	Peridotite + gabbro	40–75	9–9.8	9–19	21.0–23.3	479–485	546–549	5.9–12.9	0.064	0.13–0.28	0.25–0.43	This work
Rainbow; 36° 14' N	Peridotite + gabbro	360	2.9–3.1				>750		<2.5	2.2	13.0	19
Broken Spur; 29° N MAR	Basalt	356–360		0	11.8–12.8	419–422	469		9.30	0.06	0.43	38
Lucky Strike;	Basalt	308–324	3.8–6.4	0	32.3–36.7	347–426	417–472		2.1–3.0	0.3–0.7	0.04–0.72	20
37° 17' N MAR		185–284	3.8–3.9	0	31.3–38.2	363–428	424–514		2.0–3.0	0.4–0.8	0.003–0.27	
Menez Gwen;	Basalt	275–284	4.2–4.8	0	29.7–33.1	312–319	357–381		1.3–1.8	1.5–2.1	0.02–0.05	20
37° 50' N MAR												
Conical seamount†	Peridotite	3	9.28					30–40	2.1	0.001		11, 12
Endeavour, JdF‡	Basalt	346–370	4.2–4.5	0	13.8–42.9	260–391	350–370	0–2	3.0–8.1	1.8–3.4	0.16–0.42	39–41
21° N EPR	Basalt	273–355	3.3–3.8	0	11.7–20.8	432–510	489–579	0–0.6	6.6–8.4	0.06–0.09	0.23–1.7	40, 41
Oman Ophiolite§	Peridotite	23	11.4–11.6	0.002–0.01	1.5–1.9	11.5–35.9	9.67–26.1	0.05–0.14				42
Experiments												
	Harzburgite	300	6.4–11.6	0.002–0.02	0.29–5.24	549–576	512–541	12.1–17.8	0.6–0.8	0.066	0.10–0.33	16
	Lherzillite	200	5.4–8.0	10.7–49.4	7.5–35.7	467–500	534–560	2.04–24.8	ND	ND	ND	16
	Basalt	350	4.8	0.050	18.3	492	581	0.069	7.3	1.0	0.2	43
Theoretical	Peridotite	350	6.5–6.6	0.07–0.1	27.6–35.6	471–543	550–612	<<1.0	3.2–6.3		20.96–164.9	17

Fluids were sampled in titanium, non gas-tight samplers with Alvin. Ten millilitres of fluid was drawn into 20-ml syringes and 10-ml headspace air was added. The samples were immediately frozen at –70 °C to halt biological oxidation of gas species. Methane and H₂ values for Lost City are a minimum as gases were probably lost during sampling and/or diffusively during storage. H₂S concentrations are also minimum values because samples were not run onboard, but after ~2 weeks in cold storage. It is therefore likely that oxidation occurred. MAR, Mid-Atlantic Ridge; JdF, Juan de Fuca ridge; EPR, East Pacific Rise; ND, not detectable.

† Conical seamount is located in the Mariana forearc and contains sedimentary serpentine. It was drilled during Leg 125 of the Ocean Drilling Program. Trace element and stable isotopic compositions of carbonate chimney samples and serpentinized matrix material has been interpreted to reflect fluids with either a forearc mantle or subducted slab component, or both¹².

‡ Carbon isotopic values of δ¹³C in CH₄ of –55‰ in the Endeavour fluids are interpreted to indicate a microbial source for the methane⁴⁰.

§ Meteoric fed springs emanate from serpentinized harzburgite. The spring waters are oversaturated with respect to both serpentine and brucite. Mixing of bicarbonate-rich fluids and surface water results in precipitation of calcite or aragonite.

|| This experimental work involved reaction of harzburgitic material and an Mg-free solution at 300 °C, 500 bar, and a water-rock ratio of 10. The harzburgitic runs lasted 0–17,147 hours; additional experiments included reacting lherzillite with sea water at 200 °C, 500 bar and at water-rock ratio of 10 for 0–4,869 h.

by serpentinized peridotites (Fig. 1b). Few data are published on these three sites, but both the Rainbow and Logatchev sites host black smoker chimneys (350–360 °C)^{19–22}. High CH₄ and H₂ concentrations at these two sites indicate a peridotite influence; however, much of the chemical data (low pH values, moderate silica, Cu and Zn enrichment) are consistent with reactions involving gabbroic or basaltic material^{19,22,23}.

Recent studies suggest that hydrothermal systems similar to the LCF may be common along a significant portion of the ridge system (for example, the MAR, and the Indian and Arctic ridges). The sea-floor morphology in the vicinity of the ATF is typical of that near many large transform faults that offset the MAR and Southern Ocean ridges^{4–7}. Serpentinite bodies routinely crop out at these sites and the recovered peridotites are typically pervasively altered to serpentine minerals, indicating extensive interaction with hydrothermal fluids^{14,21,24}. Serpentinization processes have been the focus of much attention of late because of their potential importance to early Earth hydrothermal systems and because they generate significant CH₄, H₂ and possibly organic compounds during mineral–fluid reactions^{25–29}. Manifestation of such reactions is commonly inferred from CH₄ and H₂ anomalies in the water column at numerous uplifted serpentinite massifs, at highly tectonized zones believed to be peridotitic in composition, and at serpentinite outcrops along rift valley walls^{4–6,9,30–33}. A significant number of these venting sites are located on old, highly tectonized crust away from the neovolcanic zone^{4,5}. The extensive nature of these plumes suggests that such venting may play a significant role in chemical and thermal exchanges between the upper mantle and the lithosphere^{4,5,9}. However, except for the Saldanha field at 36° 30' N, located near the southern tip of the FAMOUS segment⁶, few of these sites have ever been visited.

There are many features of slow- and ultraslow-spreading systems which favour venting from off-axis environments. For example, tectonic emplacement of inside corner highs, faulting associated with transform displacements, isostatic uplift, and exfoliation induced by mass wasting probably create permeable pathways in the serpentinite basement. Within these environments, the combination of exothermic serpentinization reactions, active fracturing, and topographic forcing may drive fluid flow. Compressive stresses may also be generated on steep scarps due to the large positive volume changes (~20%) associated with serpentinization reactions. In concert, these factors promote low-temperature venting of high-pH, methane- and hydrogen-rich fluids in hydrothermal systems associated with uplifted, ultramafic massifs that are common along slow- and ultraslow-spreading ridges.

Implications for biology and early Earth hydrothermal systems

We anticipate that this newly discovered class of sea-floor hydrothermal system may provide insights into hydrothermal processes of the early Earth and the life forms that they supported^{25,34}. The reducing conditions associated with serpentinization of ultramafic material may be similar to those present in the Hadean (4.5–3.8 Gyr ago) ocean during early Earth formation and it has been suggested that such high-pH systems may have been a requirement for the emergence of life on the ocean floor^{25,35,36}. Model calculations based on thermodynamic considerations suggest that synthesis of numerous organic compounds is favoured during mixing of warm serpentinite-derived, high-pH, reducing fluids with cool, oxygenated sea water²⁵. The warm, organic- and volatile-rich environment present within the porous interior of ancient hydrothermal deposits may have been extremely suitable habitats for the emergence of thermophilic or hyperthermophilic anaerobic organisms that may represent the most ancient of lifeforms on Earth³⁷.

The LCF is an example of a previously unknown type of sea-floor chemosynthetic system that may be much more widespread than the highly localized, magmatically driven hydrothermal vent systems present along mid-ocean-ridge axes. It is a reminder of the

discoveries remaining to be made on the sea floor, which may hold important clues to the origin and diversity of life. □

Received 16 February; accepted 6 June 2001.

1. Fornari, D. J. & Embley, R. W. in *Seafloor Hydrothermal Systems, Physical, Chemical, Biological, and Geological Interactions* (eds Humphris, S. E., Zierenberg, R. A., Mullineaux, S. & Thomson, R. E.) 1–26 (American Geophysical Union Monograph 91, Washington DC, 1995).
2. Hannington, M. D., Jonasson, I. R., Herzig, P. M. & Petersen, S. in *Seafloor Hydrothermal Systems, Physical, Chemical, Biological, and Geological Interactions* (eds Humphris, S. E., Zierenberg, R. A., Mullineaux, S. & Thomson, R. E.) 115–157 (American Geophysical Union Monograph 91, Washington DC, 1995).
3. Tivey, M. K., Stakes, D. S., Cook, T. L., Hannington, M. D. & Petersen, S. A model for growth of steep-sided vent structures on the Endeavour Segment of the Juan de Fuca Ridge: Results of a petrologic and geochemical study. *J. Geophys. Res.* **104**, 22859–22883 (1999).
4. German, C. R., Parson, L. M. & HEAT Scientific Team. Hydrothermal exploration near the Azores Triple Junction: tectonic control of venting at slow-spreading ridges? *Earth Planet. Sci. Lett.* **138**, 93–104 (1996).
5. Gracia, E., Charlou, J. C., Radford-Knoery, J. & Parson, L. M. Non-transform offsets along the Mid-Atlantic Ridge south of the Azores (38° N–34° N): ultramafic exposures and hosting of hydrothermal vents. *Earth Planet. Sci. Lett.* **177**, 89–103 (2000).
6. Barriga, F. J. A. S. *et al.* Discovery of the Saldanha Hydrothermal field on the FAMOUS Segment of the MAR (36° 30' N). *Eos* **79**, 67 (1998).
7. Cann, J. R. *et al.* Correlated slip surfaces formed at ridge-transform intersections on the Mid-Atlantic Ridge. *Nature* **385**, 329–332 (1997).
8. Blackman, D. K., Cann, J. R., Janes, B. & Smith, D. K. Origin of extensional core complexes: Evidence from the Mid-Atlantic Ridge at Atlantis Fracture Zone. *J. Geophys. Res.* **103**, 21315–21333 (1998).
9. Coleman, R. G. Petrologic and geophysical nature of serpentinities. *Geol. Soc. Am. Bull.* **82**, 897–918 (1971).
10. Barnes, L., Rapp, J. R., O'Neil, J. R., Sheppard, R. A. & Gude, A. J. Metamorphic assemblages and the direction of flow of metamorphic fluids in four instances of serpentinization. *Contrib. Mineral. Petrol.* **35**, 263–276 (1972).
11. Fryer, P. *et al.* Conical Seamount: Seamount II, ALVIN submersible and seismic reflection studies. *Proc. ODP Init. Rep.* **125**, 69–94 (1990).
12. Haggerty, J. A. in *Seamounts, Islands, and Atolls* (eds Keating, B., Fryer, P. & Batiza, R.) 175–185 (American Geophysical Monograph Series 47, Washington DC, 1987).
13. Bonatti, E., Lawrence, J. R., Hamlyn, P. R. & Breger, D. Aragonite from deep sea ultramafic rocks. *Geochim. Cosmochim. Acta* **44**, 1207–1214 (1980).
14. Früh-Green, G. L., Plas, A. & Lecuyer, C. Petrologic and stable isotope constraints on hydrothermal alteration and serpentinization of the EPR shallow mantle at Hess Deep (Site 895). *Proc. ODP Sci. Res.* **147**, 255–291 (1996).
15. Sakai, R., Kusakabe, M., Noto, M. & Ishii, T. Origin of waters responsible for serpentinization of the Izu-Ogasawara-Mariana forearc seamounts in view of hydrogen and oxygen isotope ratios. *Earth Planet. Sci. Lett.* **100**, 291–303 (1990).
16. Janeky, D. R. & Seyfried, W. E. Jr Hydrothermal serpentinization of peridotite within the oceanic crust: Experimental investigations of mineralogy and major element chemistry. *Geochim. Cosmochim. Acta* **50**, 1357–1378 (1986).
17. Wetzel, L. R. & Shock, E. L. Distinguishing ultramafic from basalt-hosted submarine hydrothermal systems by comparing calculated vent fluid compositions. *J. Geophys. Res.* **105**, 8319–8340 (2000).
18. Van Dover, C. L. in *The Ecology of Deep-Sea Hydrothermal Vents* 63–69 (Princeton Univ. Press, Princeton, New Jersey, 2000).
19. Donval, J. P. *et al.* High H₂ and CH₄ content in hydrothermal fluids from Rainbow site newly sampled at 36° 14' N on the AMAR segment, Mid-Atlantic Ridge (diving FLORES cruise, July 1997). Comparison with other MAR sites. *Eos* **78**, 832 (1997).
20. Charlou, J. L. *et al.* Compared geochemical signatures and the evolution of Menez Gwen (37° 50' N) and Lucky Strike (37° 17' N) hydrothermal fluids, south of the Azores Triple Junction on the Mid-Atlantic Ridge. *Chem. Geol.* **171**, 49–75 (2000).
21. Krasnov, S. G. *et al.* in *Hydrothermal Vents and Processes* (eds Parson, L. M., Walker, C. L. & Dixon, D. R.) 43–64 (Geological Society Special Publication 87, London, 1995).
22. Douville, E., Charlou, J. L., Donval, J. P., Knoery, J. & Fouquet, Y. Trace elements in fluids from the new Rainbow hydrothermal field (36° 14' N, MAR): a comparison with other Mid-Atlantic Ridge fluids. *Eos* **78**, 832 (1997).
23. Fouquet, Y. Geological setting and compositions of hydrothermal sulfide deposits along the Mid-Atlantic Ridge. Volcanic control versus tectonic control of sulfide mineralization. *Eos* **78**, 832 (1997).
24. Lagabriele, Y. D., Bideau, D., Cannat, M., Karson, J. A. & Mevel, C. in *Faulting and Magnetism at Mid-Ocean Ridges* (eds Buck, W. R., Delaney, P. T., Karson, J. A. & Lagabriele, Y.) 153–176 (American Geophysical Union Monograph 106, Washington, DC, 1998).
25. Shock, E. L. & Schulte, M. D. Organic synthesis during fluid mixing in hydrothermal systems. *J. Geophys. Res.* **103**, 28513–28527 (1998).
26. Allen, D. A., Berndt, M. E., Seyfried, W. E. Jr & Horita, J. Inorganic reduction of CO₂ to HCOOH, CH₄, and other reduced carbon compounds with application to seafloor hydrothermal systems. *Eos* **79**, 58–59 (1998).
27. Berndt, M. E., Allen, D. E. & Seyfried, W. E. Jr Reduction of CO₂ during serpentinization of olivine at 300 °C and 500 bar. *Geology* **24**, 351–354 (1996).
28. Janeky, D. R. & Seyfried, W. E. Jr Hydrothermal serpentinization of peridotite within the oceanic crust: Experimental investigations of mineralogy and major element chemistry. *Geochim. Cosmochim. Acta* **50**, 1357–1378 (1986).
29. Neal, C. & Stanger, G. Hydrogen generation from mantle source rocks in Oman. *Earth Planet. Sci. Lett.* **66**, 315–320 (1983).
30. Karson, J. A. & Lawrence, R. M. Tectonic setting of serpentinite exposures on the western median valley wall of the MARK area in the vicinity of Site 920. *Proc. ODP Sci. Res.* **153**, 5–21 (1997).
31. Rona, P. A. *et al.* Hydrothermal circulation, serpentinization and degassing at a rift-valley fracture zone intersection: Mid-Atlantic Ridge near 15° N, 45° W. *Geology* **20**, 783–786 (1992).

32. Charlou, J. L. *et al.* Intense CH₄ degassing generated by serpentinization of ultramafic rocks at the intersection of the 15°20' N fracture zone and the Mid-Atlantic Ridge. *Geochim. Cosmochim. Acta* **62**, 2323–2333 (1998).
33. Bougault, H. *et al.* FAMOUS and AMAR segments on the Mid-Atlantic Ridge: ubiquitous hydrothermal Mn, CH₄, δ³He signals along the rift valley walls and rift offsets. *Earth Planet. Sci. Lett.* **161**, 1–17 (1998).
34. Schopf, J. W. *Earth's Earliest Biosphere: Its Origin and Evolution* (ed. Schopf, J. W.) 1–543 (Princeton Univ. Press, New Jersey, 1983).
35. Russell, M. J. & Hall, J. A. The emergence of life from iron monosulfide bubbles at a submarine hydrothermal redox and pH front. *J. Geol. Soc. Lond.* **153**, 1–26 (1996).
36. MacLeod, G., McKeown, C., Hall, H. J. & Russell, M. J. Hydrothermal and oceanic pH conditions of possible relevance to the origin of life. *Orig. Life Evol. Biosph.* **23**, 19–41 (1994).
37. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
38. James, R. H., Elderfield, H. & Palmer, M. R. The chemistry of hydrothermal fluids from the Broken Spur site, 29° N Mid-Atlantic Ridge. *Geochim. Cosmochim. Acta* **59**, 651–659 (1995).
39. Butterfield, D. A. *et al.* Gradients in the composition of hydrothermal fluids from Endeavour Ridge vent field: Phase separation and brine loss. *J. Geophys. Res.* **99**, 9561–9583 (1994).
40. Lilley, M. D. *et al.* Anomalous CH₄ and NH₄⁺ concentrations at an unsedimented mid-ocean ridge hydrothermal system. *Nature* **364**, 45–47 (1993).
41. Von Damm, K. L. in *Seafloor Hydrothermal Systems: Physical, Chemical, Biological, and Geological Interactions* (eds Humphris, S. E., Zierenberg, R. A., Mullineaux, S. & Thomson, R. E.) 222–247 (American Geophysical Union Geophysical Monograph 91, Washington DC, 1995).
42. Neal, C. & Stanger, G. Calcium and magnesium hydroxide precipitation from alkaline groundwaters in Oman, and their significance to the process of serpentinization. *Mineral. Mag.* **48**, 237–241 (1984).
43. Seewald, J. S. & Seyfried, W. E. Jr The effect of temperature on metal mobility in subseafloor hydrothermal systems: Constraints from basalt alteration experiments. *Earth Planet. Sci. Lett.* **101**, 388–403 (1990).

Acknowledgements

Shipboard party participants on cruise AT03-60 include N. Bacher, M. Basgall, D. K. Blackman, J. Cann, G. L. Frith-Green, J. S. Gee, H. Hanna, S. D. Hurst, B. E. John, J. A. Karson, D. S. Kelley, S. Lyons, J. Morgan, S. Noonan, P. Rivizzigno, D. K. Ross, G. Sasagawa and T. Schroeder. We thank the pilots, officers and crew of the RV *Atlantis-Alvin* for their professional service during this cruise. We are also grateful to the operators of *ArgoII* for their expert navigation of the camera system during the discovery exploration dive to this field. We also thank P. Hickey for piloting of *Alvin*, his sampling and his observation during the submersible dive. Support for this program was provided by the National Science Foundation.

Correspondence and requests for materials should be addressed to D.S.K. (e-mail: kelley@ocean.washington.edu).

Neurophysiological investigation of the basis of the fMRI signal

Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath & Axel Oeltermann

Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tuebingen, Germany

Functional magnetic resonance imaging (fMRI) is widely used to study the operational organization of the human brain, but the exact relationship between the measured fMRI signal and the underlying neural activity is unclear. Here we present simultaneous intracortical recordings of neural signals and fMRI responses. We compared local field potentials (LFPs), single- and multi-unit spiking activity with highly spatio-temporally resolved blood-oxygen-level-dependent (BOLD) fMRI responses from the visual cortex of monkeys. The largest magnitude changes were observed in LFPs, which at recording sites characterized by transient responses were the only signal that significantly correlated with the haemodynamic response. Linear systems analysis on a trial-by-trial basis showed that the impulse response of the neurovascular system is both animal- and site-specific, and that LFPs yield a better estimate of BOLD responses than the multi-unit responses. These findings suggest that the BOLD contrast mechanism reflects the input and intracortical processing of a given area rather than its spiking output.

Since its inception^{1–5} BOLD fMRI has generated interest not only as a tool for mapping brain activation, but also as a means of studying the dynamics of neural networks by tracking fMRI response characteristics across various spatial and temporal scales. However, BOLD responses can only measure haemodynamic changes, such as alterations in flow, blood volume, or intravascular magnetic susceptibility, leaving many unanswered questions concerning the relationship between such cerebral haemodynamic changes and actual neural activation. A number of studies in humans or animals have combined fMRI with electroencephalography (EEG)^{6–9} or optical imaging recordings of intrinsic signals¹⁰. However, optical imaging studies essentially also measure haemodynamic responses¹¹, and thus can offer little direct evidence of the underlying neural activity, whereas EEG studies typically suffer from poor spatial resolution and relatively imprecise localization of the electromagnetic field patterns associated with neural current flow.

Active states of any brain site are characterized by time-varying spatial distributions of action potentials superimposed on relatively slow-varying field potentials. An ideal combination is functional imaging in experimental animals with microelectrode recordings of such potentials, including single-spike responses and field potentials, whereby the latter relate well not only to spike activity but also to subthreshold integrative processes in areas such as dendrites that are otherwise inaccessible. Microelectrode recording methods have been used extensively to obtain a wealth of data about the central nervous system in both anaesthetized and alert animals. Adapting this technique for simultaneous electrophysiological and imaging experiments has the potential to link human studies with a large body of animal research carried out over the past 50 years; such a link may lead to answers that are unlikely to be obtained by using either technique in isolation. We present here the first of such simultaneous recordings, using a new electrophysiological measurement technique as well as new methods to actively compensate for the interference produced by the alternation of the field gradients, which are applied for image generation during echo-planar imaging (EPI) (for a comprehensive description of EPI see ref. 12).

The BOLD signal: reflection of spiking or synaptic activity?

To study the neural origin of the BOLD signal we examined the degree of correlation of the haemodynamic response to single- and multi-unit activity (MUA), as well as to LFP. Because single- and multi-unit responses were typically correlated, we will present largely MUA data in this study. MUA and LFP result from the

dynamic interaction of various synaptic and cellular mechanisms, the former reflecting primarily the output of a neural population (within approximately a couple of hundred microns of the electrode tip^{13,14}), and the latter mostly a weighted average of synchronized dendro-somatic components of the input signals of a neural population (within a few millimetres of the electrode tip^{15,16}). Thus their relationship to BOLD responses not only provides insights into the mechanisms underlying the haemodynamic changes, but it can also help us to better interpret the functional meaning of the activation patterns observed in magnetic resonance imaging.

To elicit visual cortical responses we used polar-transformed chequerboard patterns rotating at 60–180 degrees per s. The direction of rotation was reversed every 1 s to minimize adaptation. The time course of the BOLD signal was studied and correlated with the neural response by carrying out single-slice (2-mm thickness) fMRI using single- or multi-shot EPI with $0.75 \times 0.75 \text{ mm}^2$ in plane actual resolution. Optimal selection of a slice was achieved by determining the electrode position in an anatomical scan and then running a functional multi-slice (13 slices, 2-mm thickness), multi-shot EPI scan using a block design stimulation model. A new recording method, specially constructed electrodes and micro-manipulators, and a system to compensate for interference were developed to measure neural activity in the presence of the strong, alternating gradients of the magnetic field (see Methods).

Figures 1a and b show the location of the electrode tip and the BOLD response elicited by the visual stimulus in the primary visual cortex of one animal (B00). Figure 1c shows the comprehensive broadband neural signal (that is, the de-noised signal of 50 mHz to 3 kHz bandwidth, which includes LFP, MUA and single spikes) superimposed on the BOLD activation for the region of interest (ROI) around the electrode tip, that is, the average time series of the voxels inside the dotted green line. The average ROI used in the experiments was $7.1 \times 2.3 \text{ mm}^2$ (1 s.d. = 3.5 mm^2), and was defined as a small region around the tip of the electrode containing only the primary visual cortex. For any given voxel of this region the intensity of the BOLD image comprises a time series of data. Time series within this region that significantly correlated with the time course of the stimulus ($r > 0.35$; $P < 0.00001$) were selected for further analysis. The neural signal typically showed an increase of overall activity immediately on presentation of the visual stimulus that lasted for the entire duration of stimulation. The yellow trace shows the root mean square (r.m.s.) value of successive 250-ms windows.

The window size was chosen to equal the repetition time of the pulse sequence. The r.m.s. typically showed a transient, 2–3-s-long increase followed by a sustained response and then a slight decrease of activity that was also noticeable in the baseline of the de-noised raw signal. The mean onset (the time point at which the BOLD signal first increased above 3 s.d. of the baseline) of the BOLD signal was 2.16 s (1 s.d. = 0.94 s) after neural activation, reaching a plateau about 7 s later. To examine the contribution of MUA and LFP signals to the haemodynamic response we applied time-dependent frequency analysis to the raw data. Figure 1d–f shows the magnitude of the time-dependent Fourier transform (spectrogram) of the signal computed using a sliding window (Hamming window of 250 ms) for three different stimulus-presentation times (24, 12 and 4 s). Typically, after stimulus presentation a transient increase in power was observed across all frequencies, followed by a lower level of activation that was maintained during the entire stimulus presentation. A prominent characteristic in all spectrograms was a marked stimulus-induced increase in the magnitude of the LFP, which was always larger than that observed for MUA. The vertical panel across the time axis of each three-dimensional plot shows the average LFP and MUA responses computed as the mean vector of the time series between 40–130 Hz (black dashed lines) and 300–1,500 Hz (blue dashed lines) frequency regions. The vertical panel along the frequency axis shows the average spectra for the pre-stimulus, stimulation, and post-stimulus periods. A decrease in neural activity was also observed immediately after the termination of the stimulus, as shown by the deep-blue regions in all spectrograms.

To confirm these results across all scans, a spectrogram (Fig. 2) was computed for the first 6 s of the neural response averaged over all data collected during 24, 12.5, 12 and 6-s stimulus presentation.

The time series at each frequency is expressed in units of the standard deviation (s.d.) of the activity in the pre-stimulus period, and it therefore represents the signal-to-noise ratio of the response at that frequency. The maximum increase in power over all sessions (10 monkeys, $N = 619$ experiments) was found at 72.96 Hz (1 s.d. = 21.04 Hz), in other words within the gamma range of the LFPs. The top panel shows the mean LFP, MUA, total (over all frequencies) and BOLD responses for all data. There is a marked difference in signal-to-noise ratio (scale of the vertical axes) between the neural and the BOLD signals. Such a difference can, in principle, result in statistical rejection of the activation of various regions during mapping experiments, despite the fact that the underlying neural activity is highly robust and significant.

Approximately one-quarter of all multiple-unit responses were found to be transient, in that they showed an initial increase in amplitude, returning to baseline within 2–4 s. Figure 3a shows simultaneous recordings of haemodynamic responses, LFPs and transient single- and multi-unit activity in striate cortex. Both the spike-density function, representing a neuron's instantaneous firing rate, and the MUA show strong adaptation, returning to the baseline around 2.5 s after stimulus onset. In contrast, the activity underlying the LFPs remains elevated for the duration of the visual stimulus. There was no single observation period or recording site for which the opposite result was observed, namely a highly correlated MUA signal and an uncorrelated or missing LFP signal. Similarly at no time did we observe MUA that was larger in magnitude or signal-to-noise ratio than the measured LFP activity. These findings suggest that BOLD activation may actually reflect more the neural activity related to the input and the local processing in any given area, rather than the spiking activity commonly thought of as the output of the area.

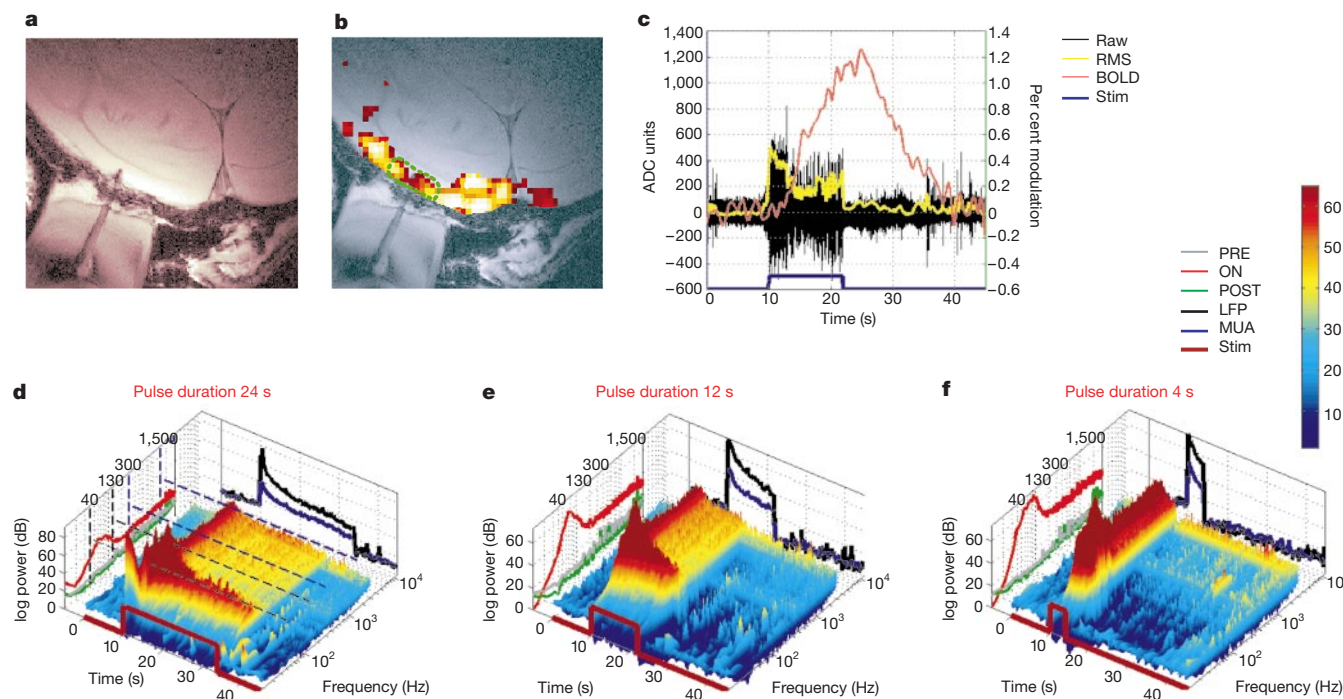


Figure 1 Neural and BOLD responses to pulse stimuli. **a**, FLASH scan (see Methods) showing the location of the electrode tip in primary visual cortex. **b**, BOLD response to rotating checkerboard patterns in striate cortex. Activation can be measured around the electrode tip. **c**, Haemodynamic response (red) superimposed on the de-noised raw neural signal (black). The term 'de-noised raw' denotes that no other signal processing beyond the removal of gradient interference (see Methods) was done. The r.m.s. of the signal is indicated by a thick yellow line. **d–f**, Spectrograms for data collected over 24, 12

and 4 s. In each three-dimensional plot, the vertical panel along the time axis shows the average LFP and MUA responses, namely the mean vector of the time series between black and blue dashed lines, respectively. The vertical panel along the frequency axis shows the average spectra for the pre-stimulus, stimulation, and post-stimulus periods. Colour bar shows the logarithm of power. ADC, Analogue to digital converter; STIM, time course of the visual stimulus; PRE, pre-stimulus period; ON, stimulus presentation period; POST, post-stimulus period.

The stronger contribution of LFPs in BOLD responses might be the result of differences in spatial summation, as LFPs usually integrate signals from a couple of millimetres, whereas MUAs do so only for a few hundred micrometres. To test whether or not such an explanation is plausible we repeated the same experiments outside the magnet, with exactly the same stimulation and anaesthesia conditions, but with a multi-unit recording system¹⁷. Intracortical recordings were carried out with a 4 × 4 array of microfibre electrodes (quartz Pt₉₀W₁₀; 80 μm shaft diameter, spacing of 250 μm centre to centre, impedance 250–750 kΩ, at 500 Hz). MUAs of distant electrodes were added using a weighting factor that decreases with the same rate as the LFPs decrease with distance from the electrode¹⁸. We (arbitrarily) assumed that MUAs are collected over an area of 200 × 200 μm², and that LFPs are collected over an area of 2,000 × 2,000 μm², which means that about 100 MUA signals (multiple recordings from each electrode) must be ‘summed’ to account for the spatial summation assumed for LFP. The contribution of MUA may even decrease by summing this type of activity. This is presumably due to the lack of significant synchronization in this frequency band. Summing the signals of a single trial slightly increased the transient portion of the MUA, but it never increased the sustained part of the response clearly seen in the LFP band (see Supplementary Information).

Finally, the transfer function of the entire recording system was carefully examined to ensure that lower frequencies were not preferentially amplified due to impedance differences at different frequencies. Both measurements and simulations showed that the gain of a frequency component monotonically increases as the frequency increases. That is, the transformation of the linear

transfer function into a constant function would attenuate MUA even more with respect to the LFP signals (see Supplementary Information).

Modelling the BOLD response from neural and imaging data

The simultaneous recording of the neural and haemodynamic responses enable the application of system-identification techniques on a trial-by-trial basis to build a model of the BOLD mechanism. We made an initial attempt to describe this mechanism by estimating its impulse response to the underlying neural activity. Correlation analysis was applied to both the measurements conducted during visual stimulation and the measurements of spontaneous activity. In either case, the input (neural) data was pre-whitened to make the results as uncorrelated as possible. Pre-whitening was done by fitting a 10th-order autoregressive model to the neural data (it corresponds to filtering ‘deviations (peaks) from whiteness’ with a filter of 60 dB oct⁻¹ attenuation and zero phase shift). The parameters of the autoregressive model were estimated by using the least-square method (that is, by minimizing the standard sum of squared, forward prediction errors). Correlation analysis was applied after subjecting the output (haemodynamic) data to the same filter. The impulse response is the cross covariance function (positive lags) of the neural and BOLD responses. Figure 4a shows the impulse functions computed from the LFP and BOLD responses from the striate cortex of two monkeys (b97, k00) under conditions of no stimulation. They were used to convolve the LFP responses to a stimulus of 100% contrast presented for 12 s (Fig. 4b) and to a stimulus with the contrast changing every 4 s (Fig. 4d). In a first approximation the model predicts the measured response quite well (session b972y1: $r^2 = 0.905$ with LFP, $r^2 = 0.725$ with MUA; session k006i1: $r^2 = 0.769$ with LFP, $r^2 = 0.625$ with MUA). However, the determination coefficient (normalized residuals) decreased rapidly with increasing stimulus duration, indicating the existence of higher-order terms ignored by linear systems analysis. Figure 4c plots the distribution of determination coefficients for BOLD estimates from LFP and MUA, using impulse

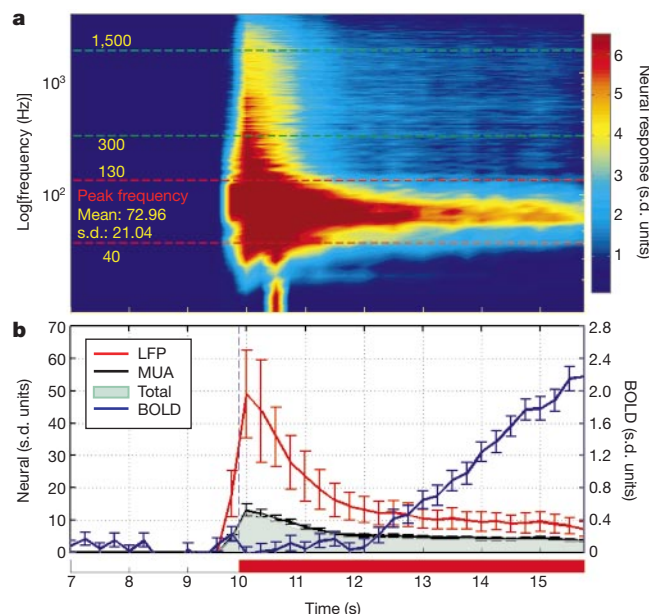


Figure 2 Time-dependent frequency analysis for population data. **a**, Spectrogram of the first 6 s of the neural response averaged over all data collected during 24, 12.5, 12 and 6 s of stimulus presentation (10 monkeys, 619 experiments). Each time course is expressed in units of the standard deviation of the pre-stimulus period. Colour (and thus the colour bar) encodes the reliability of signal change for each frequency (signal-to-noise ratio) rather than the magnitude of the power spectrum. Red and black dashed lines show the LFP and MUA frequency bands, respectively. **b**, Mean LFP (red), MUA (black) and total (green surface) neural response (average across all frequencies), together with the BOLD signal (blue). The total signal is very similar to the MUA signal as the LFP represents a small frequency range (note the logarithmic scale of the frequency axis in **a**). The figure shows the significantly higher LFP activation for both the transient and the sustained portion of the response. Error bars are 1 s.d. with $N = 10$ (number of monkeys).

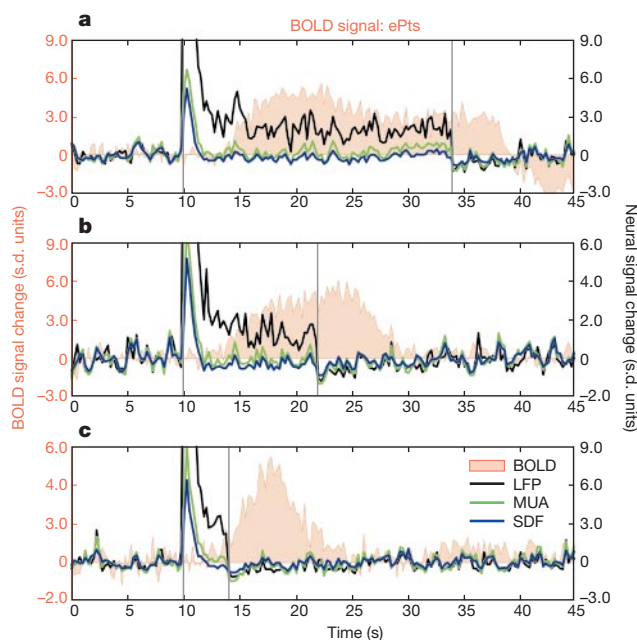


Figure 3 Simultaneous neural and haemodynamic recordings from a cortical site showing transient neural response. **a–c**, Responses to a pulse stimulus of 24, 12 and 4 s. Both single- and multi-unit responses adapt a couple of seconds after stimulus onset, with LFP remaining the only signal correlated with the BOLD response. SDF, spike-density function (see text); ePts, electrode ROI—positive time series.

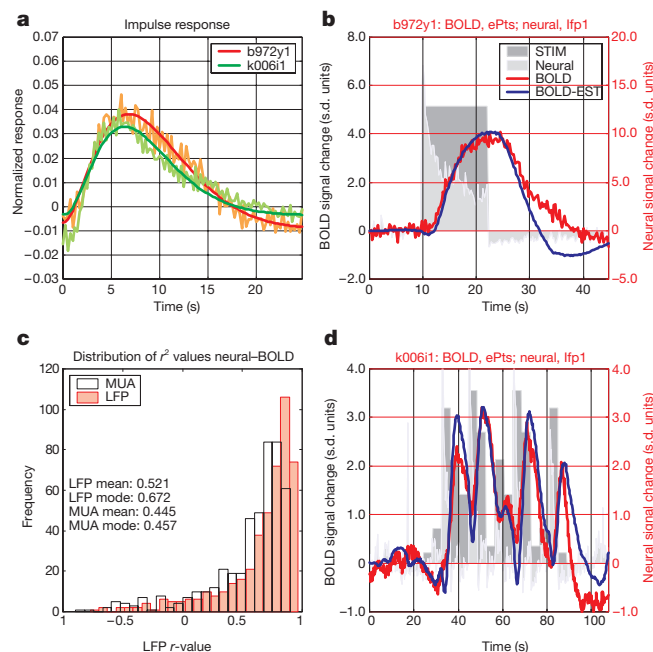


Figure 4 Correlation analysis for the estimation of the impulse response of the neurovascular system and validation of data collected with a pulse or a variable-contrast stimulus. **a**, Impulse responses for two sessions. Each computed by correlating the LFP and BOLD signal. **b, d**, Model validation. The red curve shows the measured and the blue the estimated BOLD response. The dark grey rectangle(s) show the contrast level at any given time, and the light grey curves show the neural response. **c**, Distribution of the

coefficient of determination (r^2) of LFP and MUA. The r^2 values for LFP were significantly higher than for MUA. Residual analysis (data not shown) similarly showed lower errors (in the least-square sense) in LFP-based estimates than estimates made on the basis of MUA responses. The plot in **c** includes data collected with pulse stimuli of 24, 12, 6 and 4 s ($N = 460$). ePts, electrode ROI—positive time series; lfp1, LFP from 40 to 130 Hz (see red dashed lines in Fig. 2a).

functions computed from the LFP/BOLD and MUA/BOLD covariance, respectively. On average, LFP response gave a better estimate of BOLD responses (in terms of least-square error or r^2 value) than did MUA (r_{LFP}^2 mean = 0.521, mode = 0.672; r_{MUA}^2 mean = 0.445, mode = 0.476; $P < 0.00016$, t -test).

Neural and BOLD responses as a function of stimulus contrast

To elucidate the relationship between BOLD responses and the underlying neural activity, we used four stimulus-contrast levels with four different stimulus-presentation durations to estimate the impulse response of the BOLD signal. Figure 5 shows data for a stimulus duration of 12.5 s; Fig. 5a shows fMRI responses to this stimulus. Each curve is the mean time course of the BOLD response averaged across all voxels in a ROI around the electrode tip and across all repetitions ($N = 4$). To compute the amplitude of the response a nonlinear function ($S \times \text{gamma function} \times \text{cosine}$) was fitted to the data (smooth lines in Fig. 5a). The scale S of this function was taken to be the response amplitude. Amplitudes are given as a fraction of the maximum response. As shown previously, the fMRI responses increased with stimulus contrast; this increase is directly attributable to analogously increased neural activity (Fig. 5b). Both LFP and MUA (data not shown) increased with stimulus contrast, although at a different rate than the haemodynamic responses. The thick lines in Fig. 5b were obtained by convolving the neural responses with the corresponding impulse-response function. The maximum value of these curves was taken as the neural response amplitude; it is plotted against contrast in Fig. 5c together with the BOLD responses. Neither neural nor fMRI response is a linear function of contrast, but both increase monotonically with stimulus contrast. A 'threshold' type of nonlinearity is evident for the low-contrast values, in that the fMRI response reaches about 50% of its maximum amplitude even with the lowest contrast tested (12.5%). The normalized BOLD response as a function of LFP and MUA are plotted in Fig. 5d. With the tested

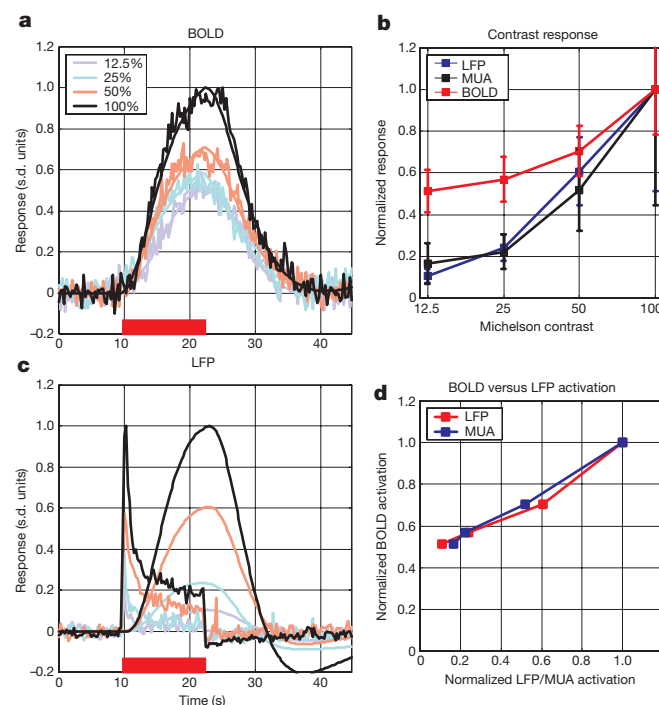


Figure 5 fMRI responses to pulse stimuli at four different contrasts (12.5, 25, 50 and 100%). **a**, Mean fMRI response superimposed with a model estimated with nonlinear curve fit. The scale parameter of the model was taken as the response amplitude. **b**, Normalized response amplitude of LFP and MUA against contrast. Data from five sessions with a pulse duration of 12.5 s. **c**, LFP responses for four different contrasts. Smooth lines are the result of convolution of the neural responses with the impulse response estimated by correlation analysis. **d**, Normalized BOLD response as a function of LFP and MUA. Responses were normalized by dividing each response by the maximum response.

contrasts, the relationship between the BOLD and the neural responses was found to be roughly linear.

Discussion

We have examined the relationship between the BOLD fMRI signal and the underlying neural activity in simultaneous intracortical electrophysiology and imaging experiments in anaesthetized monkeys. Our results show unequivocally that a spatially localized increase in the BOLD contrast directly and monotonically reflects an increase in neural activity. The time course of the haemodynamic response is roughly a low-pass-filtered expression of the total neural activity, given here as r.m.s. values in each 250-ms window of the frequency- and amplitude-modulated neural signal. Earlier studies have described the time course of the BOLD signal obtained during brief visual stimulation. After stimulus onset, the BOLD signal in humans, rats, cats or monkeys often exhibits an initial dip^{19,20}, which is attributed to the rapid increase in local deoxyhaemoglobin observed in optical imaging studies²¹. The subsequent signal increase is delayed by 2–3 s, followed by a ramp of 6–12 s to a plateau or peak value for long (>10 s) or short (<10 s) pulses, respectively. The signal returns to the baseline with a similar ramp, although often a prolonged post-stimulus undershoot is evident, which is attributed to volume changes^{20,22–24}. Our data (not presented here) show that this undershoot is often preceded by a marked inhibition in the neural response, so it may be related, at least in part, to changes in neural activity.

To better understand the neural mechanisms underlying the BOLD response, spiking and synaptic activity were examined separately by analysing single- and multi-unit activity and LFPs, respectively. LFPs are often dominated by stimulus-induced and usually stimulus-locked fast oscillations in the range of 30–150 Hz^{25–28}, as are human EEGs or magneto-encephalograms during visual or auditory tasks^{29,30}, and their role in sensory and perceptual information processing and multimodal sensory or sensory-motor integration has been debated extensively^{31,32}. Our results show that the increase in LFPs during stimulation is significantly stronger than that of multi-unit activity. Moreover, while MUA was often found to adapt, returning almost to baseline levels, LFP activity was always maintained throughout the stimulus presentation. The fast MUA decay observed in about one-quarter of the recording sites may be, at in least in part, due to lack of directional selectivity in those sites, which results in adaptation regardless of the changes in rotation direction of the stimulus. The average LFP response was also found to give better estimates of the BOLD signal (smaller error in the least-square sense) when neural activity was convolved with the neural-vascular impulse response function than when MUA was used as the system's input. In all of the measurements, the signal-to-noise ratio of the neural signal was an average of at least one order of magnitude higher than that of the fMRI signals. This observation indicates that the statistical analysis and thresholding methods applied to the haemodynamic responses probably underestimate a great deal of actual neural activity related to the stimulus or task, and suggest that a certain degree of caution is called for when interpreting mapping studies, particularly when precise localization of activity is required.

The greater contribution of LFP activity to the fMRI signal is consistent with findings regarding the bioenergetics underlying this signal. It is well established that neural activity and energy metabolism are tightly coupled³³. A quantitative relationship can actually be established between imaging signals and the cycling of certain cerebral neurotransmitters^{34–36}, as synaptic activity is tightly coupled to glucose uptake^{37,38}. NMR spectroscopy experiments exploiting such a coupling showed that the energy demands of glutamatergic neurons account for 80–90% of total cortical glucose usage in rats³⁹ and humans⁴⁰. The present findings also imply that the greater portion of the haemodynamic signal changes reflect the energetically expensive synaptic activity such as that related to the

LFP signals. Both our physiological measurements and the spectroscopy results are incompatible with models suggesting a quantitative relationship between the spike rate of neurons and the haemodynamic response⁴¹.

A number of studies have explored the contrast-response function of single cells and the relationship of the fMRI signal to the contrast. The responses of single units vary from cell to cell. In general, neurons show an increasing response that is nonlinear at low contrasts, and a linear increase up to a certain contrast level at which the response reaches its asymptote^{42,43}. Like the neural responses, fMRI BOLD response was also found to be a nonlinear function of stimulus contrast^{44,45}; however, a linear systems analysis on the fMRI responses predicted a linear relationship between the BOLD and neural activity⁴⁵. The data presented here are consistent with these studies. Furthermore they directly show that a linear relationship between the magnitude of the neural and the BOLD signals exists at the contrasts tested (12.5–100%). Nevertheless, although increases in stimulus contrast lead to monotonically increasing BOLD levels in primary visual cortex, manipulation of stimulus detectability using noise can lead to non-monotonic BOLD tuning⁴⁶.

The BOLD contrast mechanism directly reflects the neural responses elicited by a stimulus. In a first approximation BOLD and neural responses are shown to have a linear relationship for short stimulus presentation durations. Neural signals are characterized by a considerably higher signal-to-noise ratio than the haemodynamic response, suggesting that the extent of activation in human fMRI experiments is very often underestimated to a significant extent owing to the variation in the vascular response. Finally, the haemodynamic response seems to be better correlated with the LFPs, implying that activation in an area is often likely to reflect the incoming input and the local processing in a given area rather than the spiking activity. Although it is reasonable to expect that output activity will usually correlate with neurotransmitter release and pre- and post-synaptic currents, when input into a particular area primarily has a modulatory role, fMRI experiments may reveal activation in areas in which no single-unit activity is found in physiological experiments. □

Methods

Surgery, anaesthesia for fMRI and visual stimulation

This study involved 29 combined, electrophysiological fMRI experimental sessions in 10 healthy, anaesthetized monkeys (*Macaca mulatta*, 6–9 kg). All studies were carried out with great care to ensure the well being of the monkeys; were approved by the local authorities (Regierungspraesidium); and were in full compliance with the guidelines of the European Community for the care and use of laboratory animals. A detailed description of the surgical procedures has been published elsewhere²⁰ (see also Supplementary Information).

fMRI data collection

We made measurements on a vertical 4.7-T scanner with a 40-cm diameter bore (Biospec 47/40v; Bruker). The system was equipped with a 50 mTm⁻¹ (180 µs rise time) gradient coil that was actively shielded (Bruker, B-GA 26), with an inner diameter of 26 cm. A primate chair and a special transport system were designed and built for positioning the monkey within the magnet. Customized, small radio-frequency coils (30–80 mm diameter) were used (Fig. 6a), which were optimized for increased sensitivity over a given ROI, such as a portion of the primary visual cortex or regions of areas V2, V3, V4 and V5 (middle temporal visual area). The coils were attached around the recording chamber and were used as transceivers. All images were acquired using a 96 × 96 mm field of view. T1-weighted, high resolution (256 × 256 matrix, 0.5 mm thickness) anatomical scans were obtained using the three-dimensional, modified driven equilibrium Fourier transform⁴⁷ (MDEFT) pulse sequence, with an echo time of 4 ms, repetition time of 14.9 ms, flip angle of 20°, and 4 segments. To position the electrodes, 45-slice anatomical scans were acquired with the fast, low-angle shot (FLASH) sequence⁴⁸ with an echo time of 8.9 ms; repetition time of 2,000 ms; field of view of 96 × 96 mm; and a 512 × 384 matrix, reconstructed to 512 × 512, slice thickness of 0.5 mm.

Multi-slice fMRI was carried out by using multi-shot (segmented) gradient-recalled EPI (GE-EPI)⁴⁹. To examine the distribution of activation around the electrode, we initially collected volumes of 13 slices (2 mm thickness), each with a field of view of 96 × 96 mm on a 128 × 128 matrix. The acquisition parameters were: echo time of 20 ms; repetition time of 750 ms; flip angle of 40°; EPI zero phase of 8.192 ms or 40% of phase steps; pulse length

of 3.0 ms; spectral width of 100 kHz; line acquisition time of 1.28 ms; 8 segments; segment acquisition time (MRI readout window width) of 20.48 ms; repetition time between slices of 50.36 ms; and one excitation per phase-encoded step.

To study the time course of the BOLD signal and directly compare it to the neural responses, fMRI was carried out with GE-EPI with 1 or 8 interleaved segments, each segment acquired during a separate visual stimulation epoch using a 128×128 matrix, echo time/repetition time of 20/250 ms; flip angle of $20-25^\circ$; pulse length of 3.0 ms; spectral width of 100–150 kHz; EPI zero phase at 20% of phase steps; readout window width of 20.48 ms; and one excitation per phase-encoded step. The segments were

acquired separately²⁰ and merged into one image by sorting all K-space (frequency-phase-space) lines according to their phase. In these segmented acquisitions, echo time shift was used to minimize ghosting (image artefacts due to imperfections in the gradient pulses or gradient-induced eddy currents). The time of echo shift for a segment with the number j was calculated as $(1+j) \times \text{aqt} / (\text{number of EPI segments})$, where $\text{aqt} = \text{matrix size} / (2 \times \text{spectral width})$ (aqt is the readout time for a single echo). To minimize the effects of inflow and of large drainage vessels we consistently used flip angles that were smaller than the computed Ernst angle by 10° . In each session and for each monkey, an autoshim algorithm was used to tune the linear shim coils, followed by the manual tuning of the higher-order shim coils. Shimming (optimization of the homogeneity of the main magnetic field) was typically performed with a hard pulse of 50 μs duration and a receiver bandwidth of 50 kHz. Depending on the radiofrequency coil, either global or local volume shimming (FASTMAP²⁰) was used. The line width (full width at half height) at the end of the shimming procedure ranged from 30 to 70 Hz for the entire scanned volume. Together with the image acquisition both the plethysmogram (detection of volume changes due to arterial pulsations, roughly corresponding to an ECG signal) and the flow signal of the ventilator were digitized at 250 Hz, and stored for later evaluation and removal of physiological artifacts from the BOLD time courses.

MRI data analysis

All data were analysed off line on Pentium computers running the Linux or NT operating system. Analysis was performed using our own software (MATLAB) and the MEDx 3.0 image-processing package (Sensor Systems). After normalization of the data, the linear trends were removed and the data were wavelet filtered to remove temporal noise. Respiratory and cardiovascular artifacts were removed by examining the spectra of the recorded plethysmogram and respiratory signal, and creating appropriate filters. Respiration was kept constant in all experiments to 24 strokes per min, corresponding to 0.417 Hz (repetition time of 250 ms; sampling rate was 4 Hz). To increase statistical significance, all images used for three-dimensional brain reconstruction (voxel size of $1 \times 1 \times 2 \text{ mm}^3$) were spatially filtered (convolution filter with full width at half height of 1.5 mm, on a 3×3 kernel—roughly 95% of the filter is in the kernel). Both intergroup statistics (parametric paired t -test or F ratio) and correlation analysis were used to generate functional maps.

Electrophysiological recording

We developed all of the recording hardware, including electrodes, microdrives, signal conditioning and interference compensation devices (N.K.L., A.O. and M.A., manuscript in preparation). A plastic chamber (see Supplementary Information) formed to fit precisely the skull of a monkey was implanted over the occipital pole during an aseptic surgical procedure. The skull inside the chamber was left intact. At the beginning of each experiment a 2 mm trephination was performed through which the electrode was introduced into the brain. Electrodes were made of platinum-iridium ($\text{Pt}_{90}\text{Ir}_{10}$) wire etched with sodium cyanide (NaCN) solution and coated with glass (Corning glass 7570). The glass-coated tip was glued into a glass capillary tube (1.5 mm outer diameter) with super glue. The wire extended 5–10 mm beyond the end of the capillary tube and served as the contact point for the electrode. The electrode holder (Fig. 6a) consisted of three concentric, metallic cylinders (copper beryllium or a bronze alloy, CuSn_7PbZn). The innermost cylinder served as the contact point for the electrode, the middle as far-interference sensor, and the outermost layer as the amplifier ground (see below). We insulated the cylinders from each other with layers of polyetheretherketone (PEEK). The concentric cylinders, in particular the outermost cylinder, offer the additional advantage of being a rotation-symmetrical shield for the electrode, permitting optimal ground contact to the animal while avoiding any kind of loops susceptible to induction. We sealed the gap between electrode and electrode holder with silicon gel. Electrode, sensor and ground were connected to the amplifier with two twisted, coaxial cables. The outer conductor of both cables was connected to the ground cylinder of the electrode holder. The electrode holder was held firmly in a microdrive attached to the recording chamber by means of a holding ring. The positioning of the electrode was accomplished by manually turning a screw with a thread pitch of 700 μm . In addition, a three-coil, magnetic-field sensor (see below and Fig. 6a) was mounted on the microdrive. The joints between microdrive, holding ring and chamber were sealed with silicon gel. The chamber was filled with deuterium saline (0.9% NaCl in D_2O) to provide electrical contact between the ground cylinder and the animal. Deuterium saline was used instead of regular saline to avoid changing the Q-factor of the radio-frequency coil and to permit optimal shimming. Furthermore, 0.6% agar was added to the deuterium saline to minimize oscillations in both the saline and the dura during the alterations of the readout gradient. At the beginning of each session the electrode was lowered into visual cortex and positioned such that both signal intensity and stability were maximized. On the basis of magnetic resonance images most recordings were obtained from granular and infragranular layers of V1.

Signal acquisition and conditioning

Several problems arise when one tries to apply conventional neurophysiology techniques in the context of magnetic resonance imaging. Typically, a pre-amplifier is positioned near the head of the animal and is connected by means of a cable to a main amplifier some distance away. However, the strong magnetic field alternations required for the magnetic resonance image formation incapacitate the pre-amplifier by inducing voltages in any existing loop in the circuit. Shielding reduces electrical interference, but it is ineffective against magnetic interference, whereas materials such as Mu-metals that attenuate magnetic interference contain iron, which affects field homogeneity. Furthermore, low-noise, low-frequency voltage amplifiers typically used in neurophysiology rectify high-frequency voltage signals coupled with the input signal, an effect that occurs at all stages of

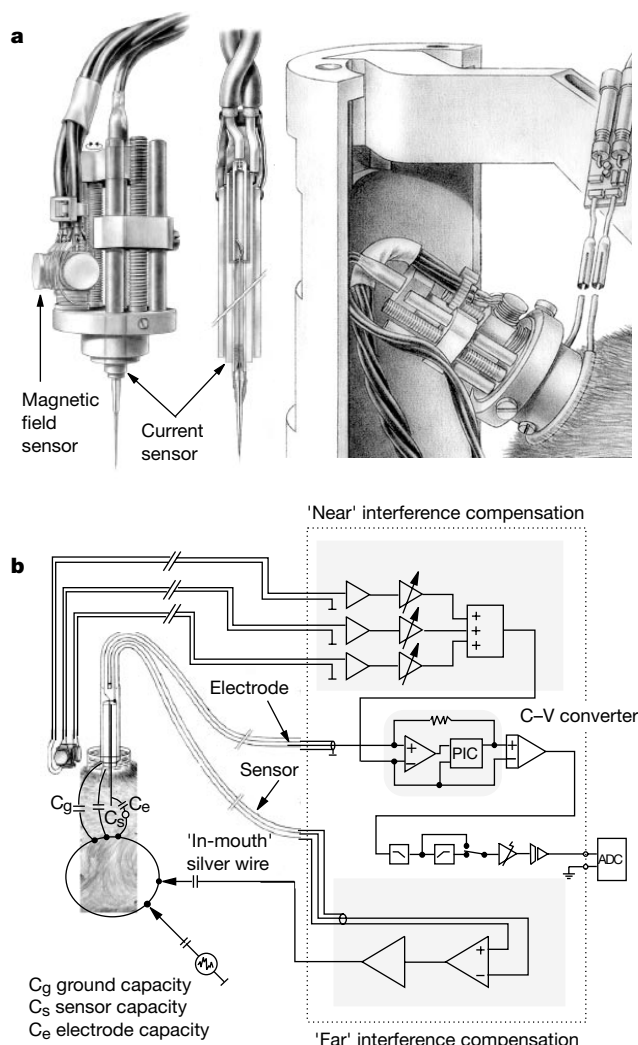


Figure 6 Recording hardware. **a**, Recordings were conducted with glass-coated, platinum-iridium electrodes. Their holder consisted of three concentric, metallic cylinders, the innermost serving as the contact point for the electrode, the middle as the far-interference sensor, and the outermost as the amplifier ground (see Methods). The cylinders were insulated from each other by layers of PEEK. Electrode, sensor and ground were connected to the amplifier with two twisted, coaxial cables (middle panel). The electrode was positioned by manually turning a screw with a fine thread pitch. In addition, a three-coil magnetic-field sensor was mounted on the microdrive to compensate for near interference (see Methods). The radio-frequency coil used for imaging was placed around the chamber. **b**, Block diagram of the recording and compensation circuitry. The animal can be conceived as being capacitively connected to any metal contact, including connections to ground (C_g), sensor (C_s) and electrode (C_e). Because of the finite animal-to-ground capacity, a fraction of the interference currents flow through the electrode. To compensate for such currents we used a sensor built into the electrode holder (current sensor). Interference originating near the electrode tip or within the electrode holder and the cables was compensated for by using three small, orthogonally oriented coils that were identical and were positioned near the electrode (magnetic field sensor). C–V, current to voltage converter; ADC, analogue to digital converter; PIC, proportional integral controller.

integrated circuits. In a 4.7-T magnet, amplifiers immediately rectify the 200 MHz radio-frequency excitation pulses, resulting in substantial disturbance of the output that only recovers several tens of milliseconds after the radiofrequency signal has subsided. Moving the conventional pre-amplifier outside of the gradient tube does not solve the problem, as the capacitance of the cable between the electrode and the pre-amplifier—in our case a cable of at least 2 m—together with the electrode impedance, act as a voltage divider compromising the signal.

To avoid signal loss owing to increased cable length, we developed a method of measuring current instead of voltage that is insensitive to cable length. Specifically, current flow from the electrode tip was measured over an ultimate cable length of 6 m and converted into voltage at the input stage of the amplifier, with pre-amplifiers, power amplifiers, operating elements and interference compensation circuits all packed into one device located outside the magnetic field. Details of the circuitry will be described elsewhere (N.K.L., A.O. and M.A., manuscript in preparation). Under our measurement conditions (roughly 600 pF input capacitance owing to cable length and high impedance of the feedback circuit of the pre-amplifier) common current-to-voltage converters are unstable. Instead of a classical current-to-voltage converter, a voltage amplifier, a proportional-integral controller (PIC) and a 10 M Ω resistor were used to ensure stability under the desired bandwidth conditions (Fig. 6b). The PIC regulates its output voltage so that the input voltage of the amplifier ultimately equals 0 V. As a result the output voltage equals the voltage drop across the 10 M Ω resistor and is therefore proportional to the measured input current. Furthermore, no signal loss occurs by charging the cable capacity. To obtain the fastest possible impulse response with the minimum possible overshoot for the desired cable length, the tuning parameters (amplification and cut-off frequency) of the Proportional Integral Controller were adjusted with a square-wave current.

Interference sources and far interference compensation

Two types of interference had to be compensated for: interference originating from a distance greater than that from the electrode tip to the electrode ground (far interference), and interference originating from the immediate vicinity of the electrode tip (near interference).

Owing to the metal–electrolyte interface the animal is capacitively connected to any metal contact (Fig. 6b), including connections to ground and electrode. Because the animal-to-ground capacity is finite, a fraction of the interference currents flowing towards the animal (for example, currents through ECG or intravenous lines) will also flow through the electrode. To compensate for such currents we used a sensor built into the electrode holder. The current between the sensor and the ground was measured, amplified and converted into a voltage. The voltage was inverted, amplified again in two steps—integrated output current limitation for the protection of the animal (10 mA maximum, 2 V maximum)—and was sent back to the animal through a junction electrode in the cavity of the mouth. This compensation set the voltage of the mouth electrode so that the current flowing through the sensor electrode remained at 0 A. In other words, the mouth current equalled the sign-inverted interference current flowing to the animal. As the three capacitances (Fig. 6b) are parallel, setting the sensor current to zero forces all the other currents to zero as well, thereby eliminating the interference measured at the tip of the electrode, and leaving the neural signal unaffected.

Near interference compensation

Interference originating and acting near the electrode tip or originating within the electrode holder and the cables cannot be detected by the sensor and compensated for by the circuit described above. The magnitude of this interference is sufficiently large to require a dedicated compensation circuit. Therefore, local changes in the magnetic field caused by the alternating field gradients during image acquisition were monitored by three small, orthogonally oriented coils that were identical and were positioned near the electrode, and the measured signal was added to the ‘ground’ of the current-to-voltage converter to neutralize the interference. This method of compensation is effective against induction currents in the electrode holder and the cable, which could not be completely avoided by the rotation-symmetrical construction described above. Furthermore this method is effective against induction currents owing to the distance between the electrode tip and ground (possibly because of eddy currents in the saline in the recording chamber and in the brain). Briefly the operation principle of this circuit is as follows. The voltages induced in the coils serve as the inputs to the compensation circuit. These are amplified and passed through an additional amplification stage, where the gain can be adjusted with a precision potentiometer to a value between -1 and 1 . The sum of the three independently adjusted signals serves as the reference (ground) for the current-to-voltage converter. The orthogonal orientation of the small coils and the adjustability of sign and value of the gain make it possible to simulate the induction voltage in a wire loop of any given diameter and orientation (vector addition). A virtual wire loop that has opposite sense of winding can be adjusted by the precision potentiometers in such a way that the remaining loops, caused by asymmetries in electrode holder and cable, are effectively compensated.

Neural data analysis

Signals were amplified by 3–30 mV μA^{-1} . In a conventional voltage-measuring system using an electrode of 300 k Ω impedance measured at 1 kHz, this would amount to a total amplification of 10^3 to 10^5 . The bandwidth of the main amplifier was 50 mHz to 3 kHz (12 dB oct^{-1} and 18 dB oct^{-1} , respectively) and the signal was digitized with 22.3 kHz using a 16-bit AD converter, set to ± 10 V input range. It was subsequently decimated by a factor of three to 7.43 kHz. From the broadband recordings three signals were extracted for analysis: (1) single-unit trains of action potentials convolved with a gaussian of fixed kernel (s.d. = 5 ms) to yield spike-density functions; (2) MUA in the range of 300–3,000 Hz; and

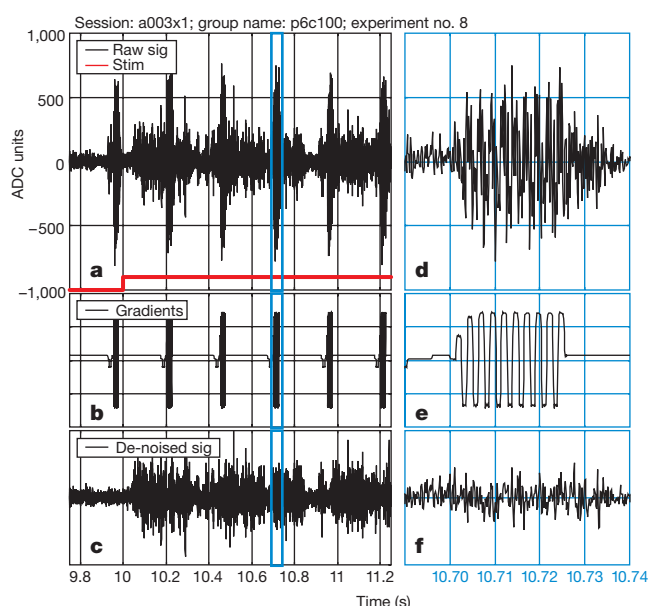


Figure 7 Elimination of residual interference by applying PCA (see Methods). **a–c**, A 1.4-s-long segment of the neural signal; **d–f**, magnification of the signal part indicated within the blue lines. The upper panels show the raw data, middle panels the recorded gradient currents, and lower ones the de-noised signal. In **a** and **d**, cell activity is shown superimposed on the strong interference induced by the gradient coils. The periodic alternations in **b** and **e** are due to the switching of the readout gradient.

(3) LFPs in the range of 10–130 Hz. Band separation was accomplished either by filtering (10–130 Hz, 36 dB oct^{-1} and 300–3,000 Hz, 48 dB oct^{-1} , both with zero phase shift) or by selectively averaging the spectrogram (generated with a Hamming window of 250 ms) of the signal across frequency. In the first case, MUA was rectified and low-pass filtered (150 Hz, 24 dB oct^{-1} , zero phase shift) to obtain the envelope of the signal. Both separation methods yielded the same results. The same results could also be obtained using conventional electrophysiological techniques (session G98; data not shown, see also Supplementary Information). Specifically, intracortical recordings were carried out with a 4×4 array of microfibre electrodes (quartz Pt₉₀W₁₀; 80 μm shaft diameter, spacing of 250 μm centre to centre, impedance 250–750 k Ω , at 500 Hz). The signal was amplified and filtered using BAK amplifiers (MMRS-1S).

Elimination of residual interference by signal processing

The techniques of interference reduction (see above) ensured a non-saturated, measurable signal, but one that was still contaminated with a certain amount of gradient interference. The elimination of the residual interference was accomplished by applying a standard mathematical, dimensionality reduction method—the principal component analysis (PCA) technique. The data were initially realigned to the slice-selection pulse (signifying the beginning of collection of an image of a single K-space segment for single- or multi-shot acquisitions) and subsequently reshaped into an N by M matrix, where N was the number of segments and M was the number of data points acquired while digitizing the physiology signal. Analysis with a personal computer of such data and elimination of those principal components that best correlated with the directly recorded interference (obtained from the current-monitor output of the gradient amplifiers) resulted in a ‘clean’ signal as shown in Fig. 7. The quality of the de-noising procedure was tested with artificial signals, and also by comparing the spectra of de-noised signals with signals acquired in the same session (interleaved) without the gradient alternations. We found no difference in their spectra.

Received 21 May; accepted 19 June 2001.

- Ogawa, S. & Lee, T. M. Magnetic resonance imaging of blood vessels at high fields: *in vivo* and *in vitro* measurements and image simulation. *Magn. Reson. Med.* **16**, 9–18 (1990).
- Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S. & Hyde, J. S. Time course EPI of human brain function during task activation. *Magn. Reson. Med.* **25**, 390–397 (1992).
- Frahm, J., Bruhn, H., Merboldt, K. D. & Hancic, W. Dynamic MR imaging of human brain oxygenation during rest and photic stimulation. *J. Magn. Reson. Imaging* **2**, 501–505 (1992).
- Menon, R. S. *et al.* Functional brain mapping using magnetic resonance imaging. Signal changes accompanying visual stimulation. *Invest. Radiol.* **27**, (Suppl.) 53 (1992).
- Kwong, K. K. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl Acad. Sci. USA* **89**, 5675–5679 (1992).
- Menon, V., Ford, J. M., Lim, K. O., Glover, G. H. & Pfefferbaum, A. Combined event-related fMRI and EEG evidence for temporal-parietal cortex activation during target detection. *NeuroReport* **8**, 3029–3037 (1997).
- Krakow, K. *et al.* EEG recording during fMRI experiments: image quality. *Hum. Brain Mapp.* **10**, 10–15 (2000).

8. Krakow, K. *et al.* EEG-triggered functional MRI of interictal epileptiform activity in patients with partial seizures. *Brain* **122**, 1679–1688 (1999).
9. Bonmassar, G., Anami, K., Ives, J. & Belliveau, J. W. Visual evoked potential (VEP) measured by simultaneous 64-channel EEG and 3T fMRI. *NeuroReport* **10**, 1893–1897 (1999).
10. Hess, A., Stiller, D., Kaulisch, T., Heil, P. & Scheich, H. New insights into the hemodynamic blood oxygenation level-dependent response through combination of functional magnetic resonance imaging and optical recording in gerbil barrel cortex. *J. Neurosci.* **20**, 3328–3338 (2000).
11. Bonhoeffer, T. & Grinvald, A. *Brain Mapping, The Methods* (eds Toga, A. W. & Mazziotta, J. C.) 55–97 (Academic, New York, 1996).
12. Schmitt, F., Stehling, M. K. & Turner, R. *Echo-Planar Imaging: Theory, Technique and Application* (Springer, Berlin, 1998).
13. Legatt, A. D., Arezzo, J. & Vaughan, H. G. J. Averaged multiple unit activity as an estimate of phasic changes in local neuronal activity: effects of volume-conducted potentials. *J. Neurosci. Methods* **2**, 203–217 (1980).
14. Freeman, W. J. *Mass Action in the Nervous System* (Academic, New York, 1975).
15. Mitzdorf, U. Properties of the evoked potential generators: current source-density analysis of visually evoked potentials in the cat cortex. *Int. J. Neurosci.* **33**, 33–59 (1987).
16. Juergens, E., Guettler, A. & Eckhorn, R. Visual stimulation elicits locked and induced gamma oscillations in monkey intracortical- and EEG-potentials, but not in human EEG. *Exp. Brain Res.* **129**, 247–259 (1999).
17. Eckhorn, R. & Thomas, U. A new method for the insertion of multiple microprobes into neural and muscular tissue, including fiber electrodes, fine wires, needles and microensors. *J. Neurosci. Methods* **49**, 175–179 (1993).
18. Juergens, E., Eckhorn, R., Fien, A. & Woelbern, T. *Brain and Evolution* 418 (Thieme, Berlin, 1996).
19. Hu, X., Le, T. H. & Ugurbil, K. Evaluation of the early response in fMRI in individual subjects using short stimulus duration. *Magn. Reson. Med.* **37**, 877–884 (1997).
20. Logothetis, N. K., Guggenberger, H., Peled, S. & Pauls, J. Functional imaging of the monkey brain. *Nature Neurosci.* **2**, 555–562 (1999).
21. Malonek, D. & Grinvald, A. Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. *Science* **272**, 551–554 (1996).
22. Buxton, R. B., Wong, E. C. & Frank, L. R. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn. Reson. Med.* **39**, 855–864 (1998).
23. Frahm, J., Kruger, G., Merboldt, K. D. & Kleinschmidt, A. Dynamic uncoupling and recoupling of perfusion and oxidative metabolism during focal brain activation in man. *Magn. Reson. Med.* **35**, 143–148 (1996).
24. Kruger, G., Kleinschmidt, A. & Frahm, J. Dynamic MRI sensitized to cerebral blood oxygenation and flow during sustained activation of human visual cortex. *Magn. Reson. Med.* **35**, 797–800 (1996).
25. Eckhorn, R. *et al.* Coherent oscillations: a mechanism of feature linking in the visual cortex? Multiple electrode and correlation analyses in the cat. *Biol. Cybern.* **60**, 121–130 (1988).
26. Murthy, V. N. & Fetz, E. E. Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys. *Proc. Natl Acad. Sci. USA* **89**, 5670–5674 (1992).
27. Gray, C. M. & Singer, W. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl Acad. Sci. USA* **86**, 1698–1702 (1989).
28. Singer, W. Synchronization of cortical activity and its putative role in information processing and learning. *Annu. Rev. Physiol.* **55**, 349–374 (1993).
29. Tallon-Baudry, C., Bertrand, O., Wienbruch, C., Ross, B. & Pantev, C. Combined EEG and MEG recordings of visual 40 Hz responses to illusory triangles in human. *NeuroReport* **8**, 1103–1107 (1997).
30. Joliot, M., Ribary, U. & Llinas, R. Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proc. Natl Acad. Sci. USA* **91**, 11748–11751 (1994).
31. Singer, W. Neuronal synchrony: a versatile code for the definition of relations? *Neuron* **24**, 49–65 (1999).
32. Shadlen, M. N. & Movshon, J. A. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* **24**, 67–77 (1999).
33. Sokoloff, L. in *Basic Neurochemistry* (eds Siegel, G., Agranoff, B., Albers, R. W. & Molinoff, P.) 565–590 (Raven, New York, 1989).
34. Magistretti, P. J., Pellerin, L., Rothman, D. L. & Shulman, R. G. Neuroscience—energy on demand. *Science* **283**, 496–497 (1999).
35. Rothman, D. L. *et al.* In vivo nuclear magnetic resonance spectroscopy studies of the relationship between the glutamate-glutamine neurotransmitter cycle and functional neuroenergetics. *Phil. Trans. R. Soc. Lond. B* **354**, 1165–1177 (1999).
36. Shulman, R. G. & Rothman, D. L. Interpreting functional imaging studies in terms of neurotransmitter cycling. *Proc. Natl Acad. Sci. USA* **95**, 11993–11998 (1998).
37. Takahashi, S., Driscoll, B. F., Law, M. J. & Sokoloff, L. Role of sodium and potassium ions in regulation of glucose metabolism in cultured astroglia. *Proc. Natl Acad. Sci. USA* **92**, 4616–4620 (1995).
38. Pellerin, L. & Magistretti, P. J. Glutamate uptake into astrocytes stimulates aerobic glycolysis: a mechanism coupling neuronal activity to glucose utilization. *Proc. Natl Acad. Sci. USA* **91**, 10625–10629 (1994).
39. Sibson, N. R. *et al.* Stoichiometric coupling of brain glucose metabolism and glutamatergic neuronal activity. *Proc. Natl Acad. Sci. USA* **95**, 316–321 (1998).
40. Pan, J. W. *et al.* Spectroscopic imaging of glutamate C4 turnover in human brain. *Magn. Reson. Med.* **44**, 673–679 (2000).
41. Rees, G., Friston, K. & Koch, C. A direct quantitative relationship between the functional properties of human and macaque V5. *Nature Neurosci.* **3**, 716–723 (2000).
42. Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17**, 8621–8644 (1997).
43. Sclar, G., Maunsell, J. H. R. & Lennie, P. Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res.* **30**, 1–11 (1990).
44. Boynton, G. M., Demb, J. B., Glover, G. H. & Heeger, D. J. Neuronal basis of contrast discrimination. *Vision Res.* **39**, 257–269 (1999).
45. Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16**, 4207–4221 (1996).
46. Rainer, G., Augath, M., Trinath, T. & Logothetis, N. K. Nonmonotonic noise tuning of BOLD fMRI signal to natural images in the visual cortex of the anesthetized monkey. *Curr. Biol.* **11**, 846–854 (2001).
47. Ugurbil, K. *et al.* Imaging at high magnetic fields: initial experiences at 4 T. *Magn. Reson. Quart.* **9**, 259–277 (1993).
48. Haase, A., Frahm, J., Matthaei, D., Hancic, W. & Merboldt, K.-D. FLASH imaging. Rapid NMR imaging using low flip-angle pulses. *J. Magn. Reson.* **67**, 258–266 (1986).
49. Mansfield, P. Multi-planar image formation using NMR spin echoes. *J. Phys. C* **10**, L55–L58 (1977).
50. Gruetter, R. Automatic, localized *in vivo* adjustment of all first- and second-order shim coils. *Magn. Reson. Med.* **29**, 804–811 (1993).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the editorial office of Nature.

Acknowledgements

We thank D. Leopold, G. Rainer and N. Sigala for reading the manuscript and for many useful suggestions. We also thank H. Mandelkow for writing some of the Matlab code; K. Lamberty for the drawings; D. Blaurock for English corrections and editing; and S. Weber for fine-mechanic work. This research was supported by the Max Planck Society.

Correspondence and requests for materials should be addressed to N.K.L. (e-mail: nikos.logothetis@tuebingen.mpg.de).

A test of general relativity from the three-dimensional orbital geometry of a binary pulsar

W. van Straten*, M. Bailes*, M. Britton*, S. R. Kulkarni†, S. B. Anderson†, R. N. Manchester‡ & J. Sarkissian‡

* Centre for Astrophysics and Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, Victoria 3122, Australia

† Division of Physics, Mathematics, and Astronomy, California Institute of Technology, Mail Code 220-47, Pasadena, California 91125, USA

‡ Australia Telescope National Facility—CSIRO, PO Box 76, Epping, New South Wales 1710, Australia

Binary pulsars provide an excellent system for testing general relativity because of their intrinsic rotational stability and the precision with which radio observations can be used to determine their orbital dynamics. Measurements of the rate of orbital decay of two pulsars have been shown^{1,2} to be consistent with the emission of gravitational waves as predicted by general relativity, but independent verification was not possible. Such verification can in principle be obtained by determining the orbital inclination in a binary pulsar system using only classical geometrical constraints. This would permit a measurement of the expected retardation of the pulse signal arising from the general relativistic curvature of space-time in the vicinity of the companion object (the ‘Shapiro delay’). Here we report high-precision radio observations of the binary millisecond pulsar PSR J0437–4715, which establish the three-dimensional structure of its orbit. We see the Shapiro delay predicted by general relativity, and we determine the mass of the neutron star and its white dwarf companion. The determination of such masses is necessary in order to understand the origin and evolution of neutron stars³.

Discovered in the Parkes 70-cm survey⁴, PSR J0437–4715 remains the closest and brightest millisecond pulsar known. It is bound to a low-mass helium white dwarf companion^{5,6} in a nearly circular orbit. Owing to its proximity, relative motion between the binary system and the Earth significantly alters the line-of-sight direction to the pulsar and, consequently, the orientation of the basis vectors used in the timing model (see Fig. 1). Although the physical orientation of the orbit in space remains constant, its parameters are measured with respect to this time-dependent basis and therefore also vary with time. Variations of the inclination angle, i , change the projection of the semi-major axis along the line-of-sight, $x \equiv a_p \sin(i)/c$, where a_p is the semi-major axis of the pulsar orbit.

The heliocentric motion of the Earth induces a periodic variation of x , known as the annual-orbital parallax⁷:

$$x^{\text{obs}}(t) = x^{\text{int}} \left[1 + \frac{\cot(i)}{d} \mathbf{r}_{\oplus}(t) \cdot \mathbf{\Omega}' \right] \quad (1)$$

The superscripts ‘obs’ and ‘int’ refer to the observed and intrinsic values, respectively, $\mathbf{r}_{\oplus}(t)$ is the position vector of the Earth with respect to the barycentre of the Solar System as a function of time, d is the distance to the pulsar, and $\mathbf{\Omega}' = \sin \Omega \mathbf{I}_0 - \cos \Omega \mathbf{J}_0$ (see Fig. 1). Similarly, the proper motion of the binary system induces secular evolution of the projected semi-major axis^{8,9}, such that:

$$\dot{x}^{\text{obs}} = \dot{x}^{\text{int}} - x \cot(i) \boldsymbol{\mu} \cdot \mathbf{\Omega}' \quad (2)$$

where $\boldsymbol{\mu} = \mu_{\alpha} \mathbf{I}_0 + \mu_{\delta} \mathbf{J}_0$ is the proper motion vector with components in right ascension, μ_{α} , and declination, μ_{δ} . An apparent transverse quadratic Doppler effect (known as the Shklovskii effect) also arises from the system’s proper motion and contributes to the observed orbital period derivative¹⁰:

$$\dot{P}_b^{\text{obs}} = \dot{P}_b^{\text{int}} + \beta P_b \quad (3)$$

where $\beta = \mu^2 d/c$, and $\mu = |\boldsymbol{\mu}|$.

Observations of PSR J0437–4715 were conducted from 11 July 1997 to 13 December 2000, using the Parkes 64 m radio telescope. Over 50 terabytes of base-band data have been recorded with the S2 Recorder¹¹ and the Caltech Parkes Swinburne Recorder (CPSR)¹², followed by offline reduction at Swinburne’s supercomputing facilities. Average pulse profiles from hour-long integrations were fitted to a high signal-to-noise template¹³, producing a total of 617 pulse time-of-arrival measurements with estimated errors of the order of 100 ns.

Previously considered negligible, the annual-orbital parallax has been largely ignored in experimental time-of-arrival analyses to date. However, our initial estimates of its peak-to-peak amplitude for PSR J0437–4715 (~ 400 ns) demonstrated that it would be clearly detectable above the timing noise. As can be seen in equation (1), x^{obs} varies with a period of one year and a phase determined by $\mathbf{\Omega}'$. Its inclusion in our timing model therefore provides a geometric constraint on Ω . We also note that the value of $\dot{x}^{\text{obs}} = (7.88 \pm 0.01) \times 10^{-14}$ observed in our preliminary studies is many orders of magnitude larger than the intrinsic $\dot{x}$ expected as a result of the emission of gravitational waves, $\dot{x}^{\text{GR}} = -1.6 \times 10^{-21}$. Neglecting $\dot{x}^{\text{int}}$, the relationship between i and Ω defined by equation (2) is parameterized by the well determined physical parameters x , $\dot{x}$ and $\boldsymbol{\mu}$. Also, because $\boldsymbol{\mu}$ is fortuitously nearly anti-parallel to $\mathbf{\Omega}'$, $\delta i / \delta \Omega$ is close to zero, and incorporation of equation (2) in our timing model provides a highly significant constraint on the inclination angle.

The orbital inclination parameterizes the shape of the Shapiro delay, that is, the delay due to the curvature of space-time about the companion. In highly inclined orbits, seen more edge-on from Earth, the companion passes closer to the line-of-sight between the pulsar and the observatory, and the effect is intensified. As the relative positions of the pulsar and companion change with binary phase, the Shapiro delay also varies and, in systems with small orbital eccentricity, is given by:

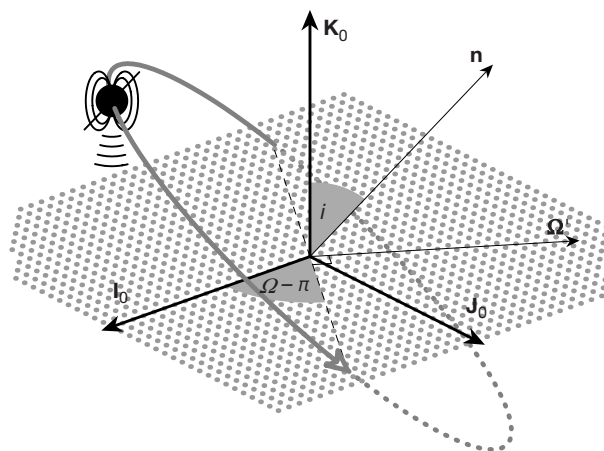


Figure 1 The three-dimensional orientation of the pulsar orbit is determined using a classical geometric model. With the centre of mass of the binary system at the origin, the basis vectors, $\mathbf{I}_0$, $\mathbf{J}_0$ and $\mathbf{K}_0$, define east, north, and the line-of-sight from Earth, respectively. The orientation of the normal vector, $\mathbf{n}$, is defined with respect to this basis by the longitude of the ascending node, Ω , and the inclination angle, i . The plane of the sky, or $\mathbf{I}_0$ – $\mathbf{J}_0$ plane (shown stippled) intersects the orbital plane at the ‘line of nodes’ (dashed line). Below the $\mathbf{I}_0$ – $\mathbf{J}_0$ plane, the orbital path has been drawn with a dotted line. The unit vector, $\mathbf{\Omega}'$, lies in the $\mathbf{I}_0$ – $\mathbf{J}_0$ plane and is perpendicular to the line of nodes. The pulsar is shown at superior conjunction, where radio pulses emitted toward Earth experience the greatest time delay owing to the gravitating mass of the companion on the opposite side of the centre of mass.

$$\Delta S = -2r \ln[1 - s \cos(\phi - \phi_0)] \quad (4)$$

Here, $s \equiv \sin(i)$ and $r \equiv Gm_2/c^2$ are known as the shape and range, respectively, ϕ is the orbital phase in radians, and ϕ_0 is the phase of superior conjunction where the pulsar is on the opposite side of the companion from Earth (as shown in Fig. 1). For small inclinations, the orbit is seen more face-on from Earth, and ΔS becomes nearly sinusoidal in form.

In the PSR J0437–4715 system, the Shapiro effect is six orders of magnitude smaller than the classical Roemer delay, the time required for light to travel across the pulsar orbit. In nearly circular orbits, the Roemer delay also varies sinusoidally with binary phase. Consequently, when modelling less inclined binary systems with small eccentricity, the Shapiro delay can be readily absorbed in the Roemer delay by variation of the classical orbital parameters, such as x . For this reason, a previous attempt at measuring the Shapiro effect in the PSR J1713+0747 system¹⁴ yielded only weak, one-sided limits on its shape and range.

In contrast, we have significantly constrained the shape independently of general relativity, enabling calculation of the component of ΔS that remains unabsorbed by the Roemer delay. The theoretical signature is plotted in Fig. 2 against post-fit residuals obtained after fitting the time-of-arrival data to a model that omits the Shapiro effect. To our knowledge, this verification of the predicted space-time distortion near the companion is the first such confirmation (outside our Solar System) in which the orbital inclination was determined independently of general relativity.

The range of the Shapiro delay provides an estimate of the companion mass, $m_2 = 0.236 \pm 0.017 M_\odot$, where $M_\odot$ is the mass of the Sun. Through the mass function³, $f(M)$, we then obtain a measurement of the pulsar mass $m_p = 1.58 \pm 0.18 M_\odot$. Slightly heavier than the proposed average neutron star mass³, $m_p^{\text{avg}} = 1.35 \pm 0.04 M_\odot$, this value of m_p suggests an evolutionary scenario that includes an extended period of mass and angular momentum transfer. Such accretion is believed to be necessary for a neutron star to attain a spin period of the order of a millisecond¹⁹. It is also expected that, during accretion, the pulsar spin and orbital angular momentum vectors are aligned. Under this assumption, the measured inclination angle of $i = 42.75 \pm 0.09^\circ$ does not support the conjecture that pulsar radiation may be preferentially beamed in the equatorial plane¹⁵.

The total system mass, M , can be calculated from the observed $\dot{\omega}$,

Table 1 PSR J0437–4715 physical parameters

Right ascension, α (J2000)	04 h 37 m 15.7865145(7) s
Declination, δ (J2000)	−47° 15′ 08.461584(8)″
μ_α (mas yr ^{−1})	121.438(6)
μ_δ (mas yr ^{−1})	−71.438(7)
Annual parallax, π (mas)	7.19(14)
Pulse period, P (ms)	5.757451831072007(8)
Reference epoch (MJD)	51194.0
Period derivative, $\dot{P}$ (10 ^{−20})	5.72906(5)
Orbital period, P_b (days)	5.741046(3)
x (s)	3.36669157(14)
Orbital eccentricity, e	0.00019186(5)
Epoch of periastron, T_0 (MJD)	51,194.6239(8)
Longitude of periastron, ω (degrees)	1.20(5)
Longitude of ascension, Ω (degrees)	238(4)
Orbital inclination, i (degrees)	42.75(9)
Companion mass, m_2 ($\times M_\odot$)	0.236(17)
$\dot{P}_b$ ($\times 10^{-12}$)	3.64(20)
$\dot{\omega}$ (degrees yr ^{−1})	0.016(10)

Best-fit physical parameters and their formal 1σ errors were derived from arrival time data by minimizing an objective function, χ^2 , as implemented in TEMPO (<http://pulsar.princeton.edu/tempo>). Our timing model is based on a relativistic binary model¹⁸ and incorporates additional geometric constraints derived by Kopeikin². Indicative of the solution's validity, χ^2 was reduced by 30% with the addition of only one new parameter, Ω . To determine the 1σ confidence intervals of Ω and i , we mapped projections of the $\Delta\chi^2 \equiv \chi^2(\Omega, i) - \chi^2_{\text{min}} = 1$ contour, where $\chi^2(\Omega, i)$ is the value of χ^2 minimized by variation of the remaining model parameters, given constant Ω and i . Parenthesized numbers represent uncertainty in the last digits quoted, and epochs are specified using the Modified Julian Day (MJD).

using the general relativistic prediction of the rate of orbital precession. Using M , $f(M)$, and i , we obtain a second consistent estimate of the companion mass, $m'_2 = 0.23 \pm 0.14 M_\odot$, the precision of which is expected to increase with time as $t^{3/2}$, surpassing that of the r -derived value in approximately 30 years.

The complete list of physical parameters modelled in our analysis is included in Table 1. Most notably, the pulsar position, parallax distance, $d_\pi = 139 \pm 3$ pc, and proper motion, $\mu = 140.892 \pm 0.006$ mas yr^{−1}, are known to accuracies unsurpassed in astrometry. Although closer, d_π lies within the 1.5σ error of an earlier measurement⁹, 176 ± 26 pc. The d_π and μ estimates can be used to calculate β and the intrinsic spin period derivative, $\dot{P}^{\text{int}} = \dot{P}^{\text{obs}} - \beta P = (1.86 \pm 0.08) \times 10^{-20}$, providing an improved characteristic age of the pulsar, $\tau_c = P/(2\dot{P}^{\text{int}}) = 4.9$ Gyr. Another distance estimate may be calculated using the observed μ and $\dot{P}_b$ by solving equation (3) for d , after noting the relative negligibility of any intrinsic contribution¹⁶. The precision of the derived value, $d_b = 150 \pm 9$ pc, is anticipated to improve as $t^{5/2}$,

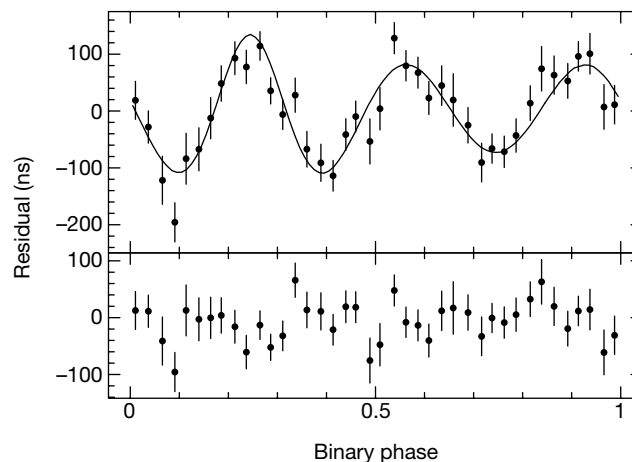


Figure 2 Arrival time residuals confirm the predicted space-time distortion induced by the pulsar companion. The unabsorbed remnant of the Shapiro delay is much smaller than the theoretical total delay, which for PSR J0437–4715 has a peak-to-peak amplitude of about 3.8 μ s. In the top panel, the solid line models the expected delay resulting from a companion with a mass of $0.236 M_\odot$, at the geometrically determined orbital inclination.

Measured arrival time residuals, averaged in 40 binary phase bins and plotted with their 1σ errors, clearly exhibit the predicted signature. In the bottom panel, the same residuals with the model removed have an r.m.s. residual of only 35 ns and a reduced χ^2 of 1.13.

providing an independent distance estimate with a relative error of about 1% within the next three to four years.

With a post-fit residual root mean square (r.m.s.) of just 130 ns over 40 months, the accuracy of our analysis has enabled the detection of annual-orbital parallax. This has yielded a three-dimensional description of a pulsar binary system and a new geometric verification of the general relativistic Shapiro delay. Only the Space Interferometry Mission (SIM) is expected to localize celestial objects with a precision similar to that obtained for PSR J0437–4715 (including parallax). By the time SIM is launched in 2010, the precision of this pulsar's astrometric and orbital parameters will be vastly improved. Observations of the companion of PSR J0437–4715 using SIM will provide an independent validation and a tie between the SIM frame and the Solar System dynamic reference frame.

We also expect that continued observation and study of this pulsar will ultimately have an important impact in cosmology. Various statistical procedures have been applied to the unmodelled residuals of PSR B1855+09 (see ref. 17 and references therein) in an effort to place a rigorous upper limit on Ω_g , the fractional energy density per logarithmic frequency interval of the primordial gravitational wave background. As the timing baseline for PSR J0437–4715 increases, our experiment will probe more deeply into the low frequencies of the cosmic gravitational wave spectrum, where, owing to its steep power-law dependence¹⁷, the most stringent restrictions on Ω_g can be made. □

Received 22 March; accepted 29 May 2001.

1. Taylor, J. H. & Weisberg, J. M. A new test of general relativity: Gravitational radiation and the binary pulsar PSR 1913+16. *Astrophys. J.* **253**, 908–920 (1982).
2. Stairs, I. H. *et al.* Measurement of relativistic orbital decay in the PSR B1534+12 binary system. *Astrophys. J.* **505**, 352–357 (1998).
3. Thorsett, S. E. & Chakrabarty, D. Neutron star mass measurements. I. Radio pulsars. *Astrophys. J.* **512**, 288–299 (1999).
4. Johnston, S. *et al.* Discovery of a very bright, nearby binary millisecond pulsar. *Nature* **361**, 613–615 (1993).
5. Bell, J. F., Bailes, M. & Bessell, M. S. Optical detection of the companion of the millisecond pulsar PSR J0437–4715. *Nature* **364**, 603–605 (1993).
6. Danziger, I. J., Baade, D. & Della Valle, M. Optical spectroscopy and photometry of the companion of the bright millisecond pulsar PSR J0437–4715. *Astron. Astrophys.* **276**, 382–388 (1993).
7. Kopeikin, S. M. On possible implications of orbital parallaxes of wide orbit binary pulsars and their measurability. *Astrophys. J.* **439**, L5–L8 (1995).
8. Kopeikin, S. M. Proper motion of binary pulsars as a source of secular variation of orbital parameters. *Astrophys. J.* **467**, L93–L95 (1996).
9. Sandhu, J. S. *et al.* The proper motion and parallax of PSR J0437–4715. *Astrophys. J.* **478**, L95–L98 (1997).
10. Damour, T. & Taylor, J. H. On the orbital period change of the binary pulsar PSR 1913+16. *Astrophys. J.* **366**, 501–511 (1991).
11. Cannon, W. H. *et al.* The S2 VLBI system. *Vistas Astron.* **41**, 297–302 (1997).
12. van Straten, W., Bailes, M. & Britton, M. In *Pulsar Astronomy—2000 and Beyond*, IAU Colloq. 177 (eds Kramer, M., Wex, N. & Wielebinski, R.) 283–284 (Astron. Soc. Pacif. Conf. Series, San Francisco, 2000).
13. Taylor, J. H. Pulsar timing and relativistic gravity. *Proc. R. Soc. Lond. A* **341**, 117–134 (1992).
14. Camilo, F., Foster, R. S. & Wolszczan, A. High-precision timing of PSR J1713+0747: Shapiro delay. *Astrophys. J.* **437**, L39–L42 (1994).
15. Backer, D. C. The neutron star-helium white dwarf population in the galactic disc. *Astrophys. J.* **493**, 873–878 (1998).
16. Bell, J. F. & Bailes, M. A new method for obtaining binary pulsar distances and its implications for tests of general relativity. *Astrophys. J.* **456**, L33–L36 (1996).
17. McHugh, M. P., Zalamansky, G., Vernott, F. & Lantz, E. Pulsar timing and the upper limits on a gravitational wave background: A bayesian approach. *Phys. Rev. D* **54**, 5993–6000 (1996).
18. Damour, T. & Deruelle, N. General relativistic celestial mechanics of binary systems. II. The post-Newtonian timing formula. *Ann. Inst. H. Poincaré Phys. Théor.* **44**, 263–292 (1986).
19. Taam, R. E. & van den Heuvel, E. P. J. Magnetic field decay and the origin of neutron star binaries. *Astrophys. J.* **305**, 235–245 (1986).

Acknowledgements

The Parkes Observatory is part of the Australia Telescope which is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO. We thank the staff at Parkes Observatory for technical assistance and performance of regular observations. S.R.K. and S.B.A. thank NSF and NASA for supporting their work at Parkes. We also thank R. Edwards and M. Toscano for comments on the text. We received support from Compaq and the Space Geodynamics Laboratory of the Centre for Research in Earth and Space Technology. M. Bailes is an ARC Senior Research Fellow.

Correspondence and requests for materials should be addressed to W.v.S. (e-mail: wvanstra@mania.physics.swin.edu.au).

Discovery of water vapour around IRC+10216 as evidence for comets orbiting another star

Gary J. Melnick*, David A. Neufeld†, K. E. Saavik Ford†, David J. Hollenbach‡ & Matthew L. N. Ashby*

*Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

†Department of Physics & Astronomy, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland 21218, USA

‡NASA/Ames Research Center, Moffett Field, California 94035, USA

Since 1995, planets with masses comparable to that of Jupiter have been discovered around approximately 60 stars¹. These planets have not been seen directly, but their presence has been inferred from the small reflex motions that they gravitationally induce on the star they orbit; these motions result in small periodic wavelength shifts in the stellar spectrum. The presence of analogues of the smaller bodies in our Solar System cannot, however, be determined using this technique, because the induced reflex motions are too small—so an alternative approach is needed. Here we report the observation of circumstellar water vapour around the ageing carbon star IRC+10216; water is not expected in measurable quantities around such a star. The only plausible explanation for this water is that the recent evolution of IRC+10216, which has been accompanied by a prodigious increase in its luminosity, is causing the vaporization of a collection of orbiting icy bodies—a process considered in an earlier theoretical study².

The notion that large numbers of icy bodies may orbit other stars has its parallel in the Solar System's Kuiper belt. The Kuiper belt is a collection of cometary nuclei, located roughly in the plane of the ecliptic and beyond the orbit of Neptune, beginning at approximately 30 astronomical units (AU; 1 AU = 1.5×10^{13} cm) from the Sun, and thought possibly to extend to a few hundred AU from the Sun. About 10^5 Kuiper-belt objects exist with diameters larger than 100 km—the largest being Pluto ($\sim 2,400$ km)—and it is likely that there is a significantly greater number of smaller objects. Short-period comets have orbital characteristics that suggest that their present-day source is the Kuiper belt. Long-period comets are presumed to originate from the Oort cloud, an approximately spherically symmetric cloud of cometary nuclei located well beyond the Kuiper belt at distances from the Sun of between 3,000 and 10^5 AU. For main-sequence stars, such as the Sun, the stellar radiation field is sufficiently weak beyond about 3 AU that icy bodies are unaffected. Even stars possessing masses between 1.5 and 4 times that of the Sun, the range which is believed to encompass the mass of IRC+10216, would have lacked the luminosity to have vaporized icy bodies beyond about 40 AU during their main-sequence phase.

After exhausting the supply of hydrogen and helium in its core, a star of a few solar masses evolves off the main sequence and begins burning hydrogen and helium in thick shells surrounding a core enriched in carbon. By the time the star has reached this so-called asymptotic giant branch (AGB) phase, the stellar radius has increased by a factor of several hundred to a thousand—in the case of IRC+10216 reaching a value of about 5 AU (the radius of Jupiter's orbit)—while the luminosity has increased by a factor of between 100 and 3,000. This luminosity increase is sufficient to cause the vaporization of any icy body orbiting within several hundred AU (ref. 2), thus resulting in the release of significant amounts of water vapour—provided that the star is surrounded by a Kuiper-belt analogue. Objects orbiting in an analogue of the Oort cloud, however, would remain unaffected.

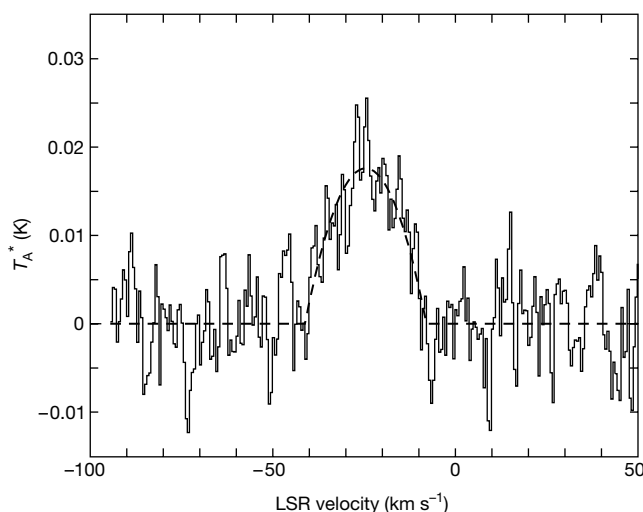


Figure 1 The spectrum of the $1_{10}-1_{01}$ transition of *ortho*-H₂¹⁶O obtained from IRC+10216. This spectrum, at 556.936 GHz, was obtained by the SWAS satellite: the continuum has been subtracted. The dashed line is a parabolic curve fitted to the spectrum (see text). All data were obtained with the telescope pointed at the position $\alpha = 9\text{ h }47\text{ min }57.4\text{ s}$, $\delta = 13^\circ 16' 44''$ (J2000). The observations were performed by 'nodding' the observatory between the source position and a reference position 30 arcmin north of the source, selected to coincide with a region of no detectable molecular emission. The SWAS field of view at 557 GHz is 3.3×4.5 arcmin. A total of 196.4 hours of on-source integration time was obtained, using three different local oscillator settings. The water line was evident in the signal sideband at all three settings. The data were reduced using the standard SWAS pipeline. Analyses of the spectral noise at frequencies separated from the expected line confirm that the system performs in accordance with radiometer theory (that is, the signal-to-noise ratio improves in proportion to $\sqrt{\text{time}}$) and that the noise obeys gaussian statistics¹⁸. LSR, local standard of rest; T_A^* , antenna temperature.

Unfortunately, the detection of water vapour in the spectrum of an evolved star does not unambiguously constitute evidence for the presence of Kuiper-belt-type objects. For those cases in which the circumstellar gas is oxygen-rich, it is predicted that almost all of the carbon will combine with oxygen to form CO, with much of the remaining oxygen reacting to form H₂O. Strong water-vapour emission is expected from such oxygen-rich stars and, indeed, this is observed (see, for example, refs 3 and 4). During the final phase of their evolution, however, convection of material from the core can enhance the outer layers of AGB stars with carbon, altering the carbon-to-oxygen ratio in the outflowing gas. For those AGB stars, like IRC+10216, possessing circumstellar envelopes in which carbon is the most abundant heavy element, it is predicted—by contrast—

that the equilibrium chemistry will drive all of the oxygen into CO with little remaining to form other molecules⁵. Thus, the detection of water vapour toward a carbon-rich AGB star raises the possibility that icy bodies are being vaporized, since the carbon-rich outflow from the star has no chemical route to produce significant water vapour.

Of the (more than 100) stars that are known to have carbon-rich circumstellar envelopes, IRC+10216 is by far the brightest and most intensively studied. We conducted a search for water vapour surrounding IRC+10216 in the rotational ground-state $1_{10}-1_{01}$ 556.936-GHz line of *ortho*-H₂¹⁶O using the Submillimeter Wave Astronomy Satellite (SWAS) during the periods 11–31 May 1999, 12 November–12 December 1999, 24 December 1999–8 January 2000, and 5 February–9 March 2000. The spectrum obtained is shown in Fig. 1. The line profile is well fitted by the parabolic curve expected for optically thick emission from an unresolved, constant-velocity outflow. The integrated antenna temperature is 0.39 K km s^{-1} , uncorrected for the SWAS aperture efficiency of 0.66. The line width at half-maximum is 25.0 km s^{-1} and the line centre is at -23.5 km s^{-1} (with respect to the local standard of rest), in good agreement with the dynamical quantities measured in the lines of other species toward this source (see, for example, ref. 6).

An identification other than the $1_{10}-1_{01}$ water-vapour line can be ruled out. Standard line catalogues⁷, although admittedly incomplete, reveal no molecules possessing both a transition with a frequency within a few km s^{-1} of 556.936 GHz and an upper-level fractional population that is plausibly high enough to account for the observed line strength. Furthermore, due to its low excitation energy and to the small partition function of the water molecule, the $1_{10}-1_{01}$ water transition is excited very effectively and thus the abundance required of any other molecule responsible for the observed feature would have to be considerably greater than that derived below for water. The only plausible molecule of sufficient abundance is HCN, for which a complete line catalogue⁸ reveals no transition of the required frequency in any vibrational state of energy lower than 15,000 K.

The water abundance is derived by modelling the circumstellar envelope as an outflow with the properties given in Table 1. Over the range of mass-loss rates inferred previously^{9,10} for IRC+10216, $\dot{M} = (2-5) \times 10^{-5} M_\odot \text{ yr}^{-1}$, we find that the $1_{10}-1_{01}$ line luminosity (in W) can be expressed as

$$L(1_{10}-1_{01}) = 3.6 \times 10^{21} (\dot{M}/10^{-5}) (x[\text{H}_2\text{O}]/10^{-7})^{0.5}$$

where $x[\text{H}_2\text{O}]$ is the water abundance relative to H₂, and $\dot{M}$ is in units of $M_\odot \text{ yr}^{-1}$. The $1_{10}-1_{01}$ line flux measured by SWAS is $1.03 \times 10^{-20} \text{ W cm}^{-2}$, corresponding to a luminosity of $3.6 \times 10^{22} \text{ W}$ for an isotropic source at an assumed distance of 170 pc (ref. 9). The implied water abundance is:

$$x[\text{H}_2\text{O}] = 1.1 \times 10^{-6} [\dot{M}/(3 \times 10^{-5})]^{-2.0} = (4-24) \times 10^{-7}$$

Table 1 Assumptions in the model for IRC+10216 circumstellar outflow

Physical conditions	Value	Reference
Stellar radius	$7.65 \times 10^{13} \text{ cm}$	9
Photospheric temperature	2,010 K	9
Dust-shell inner radius, R_i	$2.58 \times 10^{14} \text{ cm}$	9
Dust-shell outer radius	$1 \times 10^{17} \text{ cm}$	9
Outflow velocity, v_r	$14.5 [1.00 - 0.95(R/R_i)]^{0.5} \text{ km s}^{-1}$	9
Turbulent velocity	0.65 km s^{-1}	9
H ₂ density*	$(3.11 \times 10^7 \text{ cm}^{-3}/R_{15}^2) \times (\dot{M}/3 \times 10^{-5} M_\odot \text{ yr}^{-1}) \times (14.5 \text{ km s}^{-1}/v_r)$	10
Gas temperature	$\text{Max}[10, 12(90/R_{15})^{0.72}] \text{ K}$	10
Dust temperature	$1,300 (R/R_i)^{-0.4} \text{ K}$	
Opacity due to dust	$0.5 \text{ cm}^2 (\text{wavelength}/50 \mu\text{m})^{-1.3} \text{ per gram of gas}$	
Assumed distance	170 pc	9

Our model yields predictions for the luminosities of several water transitions as a function of water abundance, using two independent methods that are found to be in excellent agreement with each other: a fast escape probability method based upon the assumption that the large velocity gradient approximation applies; and a more exact Monte Carlo method^{16,17} that is much more expensive numerically. The effects of radiative pumping of rotational transitions by the far-infrared radiation field are included; vibrational excitation effects are negligible.

* In the second column, R_{15} is the radial distance (in cm) divided by 10^{15} , R is the radius within the outflow, and $\text{Max}[x,y]$ means that this quantity assumes the larger of the two values x and y .

The derived water abundance in IRC+10216 can be compared with the abundance that would be expected if orbiting icy bodies are not present. In thermochemical equilibrium, the initial water abundance in the outflowing gas is expected to be only $\sim 10^{-12}$ (ref. 11). The chemical composition of the gas can subsequently be altered by shocks¹¹ in the inner part of the circumstellar envelope—a process invoked previously to explain observations of water vapour towards the protoplanetary nebula CRL618 (ref. 12)—or by photodissociation in the outer envelope⁵. The best current models^{5,11} for these two processes, constructed specifically for the source IRC+10216, indicate (1) that pulsationally-driven shocks—

if present—would actually reduce the water abundance below its already small thermochemical equilibrium value¹¹; and (2) that photodissociation of CO followed by incorporation of O into H₂O can indeed increase the water-vapour abundance, but only to a value of about 5×10^{-12} (T. J. Millar and E. Herbst, personal communication). Thus, apart from the evaporation of icy bodies, there is no known mechanism that comes within a factor of 10^4 of explaining the water abundance observed in IRC+10216.

We now turn to the model for evaporation, in which the evaporation of orbiting icy bodies is driven by the increasing luminosity of IRC+10216. The stellar evolution is characterized by a rise in luminosity, modulated by periodic stellar thermal pulsations. As the luminosity of the star increases, a zone of evaporation moves outwards through the orbiting material (Fig. 2). The inner edge of the evaporation zone is defined by the distance within which the largest icy bodies have already been entirely vaporized, and the outer edge is defined by the distance at which the icy bodies are warmed above the sublimation temperature of water ice. By the time the convection of carbon has made the photospheric C/O ratio greater than 1, it is found that for distances up to 75 AU even icy bodies as large as Pluto will have been completely vaporized (K.E.S.F. and D.A.N., manuscript in preparation); thus, in the present evaporation model, the release of water vapour currently observed occurs beyond about 75 AU. The evaporation is a continuous process, in which the continuing vaporization of icy bodies replenishes the water vapour as it flows through the outflow and is ultimately photodissociated by the interstellar ultraviolet radiation field. New icy bodies continuously replenish the sublimed objects because the evaporation zone steadily moves outwards as the stellar luminosity increases.

Detailed models for the evaporation of icy bodies have been constructed by K.E.S.F. and D.A.N. (manuscript in preparation), but a simple argument reproduces the essential result. The current mass loss rate of water (in units of $M_{\odot} \text{ yr}^{-1}$) is given by

$$\dot{M}x(\text{H}_2\text{O})[\mu(\text{H}_2\text{O})/\mu(\text{H}_2)] \approx 3 \times 10^{-10}$$

where $[\mu(\text{H}_2\text{O})/\mu(\text{H}_2)] = 9$ is the ratio of the molecular mass of water to that of H₂. As the high current rate of hydrogen mass-loss can have been sustained for only 10^5 yr, a total ice mass of $\sim 3 \times 10^{-5} M_{\odot} \approx 10$ Earth masses is sufficient to supply the necessary water vapour. This is comparable to the original mass of water ice believed to have been present in the Solar System's Kuiper belt¹³.

The range of radii over which water vapour is present in the outflow, and therefore the range of excitation conditions, determines the emergent water spectrum. A complete (2.4–197 μm) spectrum of IRC+10216, obtained by the Infrared Space Observatory (ISO)^{14,15}, allows upper limits to be placed on a number of water transitions. Based upon our line flux predictions, the most easily detectable H₂O line in this wavelength region should be the $2_{12}-1_{01}$ 179.53- μm transition. Unfortunately, because this line is blended with a stronger HCN line at 179.51 μm , a flux determination in the water line is difficult. The next most easily detectable water line in this band is predicted to be the $3_{03}-2_{12}$ transition at 174.63 μm , which is free from overlap with other lines. The upper limit on the ISO-measured strength of the H₂O $3_{03}-2_{12}$ line is marginally below what would be expected were the water abundance uniform throughout the circumstellar envelope, but is entirely consistent with our model—in which the water emission arises from the cooler (~ 100 K), lower-density gas beyond 75 AU where icy bodies are evaporating. □

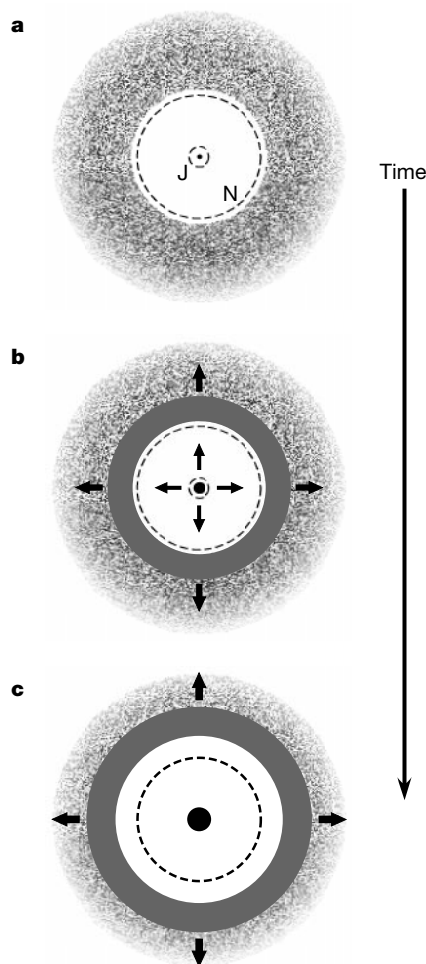


Figure 2 Schematic drawing of the evolution of IRC+10216. The star evolves from its main sequence to its asymptotic giant branch phase accompanied by an increase in both its size and luminosity, causing a wave of evaporation to propagate through the orbiting icy bodies. **a**, During its main-sequence phase, the luminosity of IRC+10216 is insufficient to vaporize icy bodies orbiting more distant than a few AU—for scale, distances equal to the orbits of Jupiter (J) and Neptune (N) are shown as dashed circles. **b**, With the hydrogen and helium depleted within its core, IRC+10216 begins to expand, its luminosity increases, and a zone of vaporization (solid grey) begins to move outward through the population of icy bodies, depositing water vapour into the outflow. **c**, The diameter of IRC+10216 has increased to a size roughly equivalent to the orbit of Jupiter, its luminosity has increased greatly, and the zone of vaporization expands. Detailed models by K.E.S.F. and D.A.N. (manuscript in preparation) describe the evolution of a collection of icy bodies with distributions in size and orbital semi-major axis that are similar to those characterizing the Solar System's Kuiper belt¹³. They indicate that water abundances as large as $\sim 10^{-7} (M_0/M_{\oplus})$ can be achieved in the outflow, where M_0 is the mass of orbiting water ice at the onset of the post-main-sequence evolution and $M_{\oplus}$ is the mass of the Earth. This result is reproduced by a simple argument given above in the text.

Received 26 February; accepted 4 June 2001.

1. Marcy, G. W. & Butler, P. R. Planets orbiting other stars. *Publ. Astron. Soc. Pacif.* **112**, 137–140 (2000).
2. Stern, S. A., Shull, M. J. & Brandt, J. C. Evolution and detectability of comet clouds during post-main-sequence stellar evolution. *Nature* **345**, 305–308 (1990).
3. Woolf, N. J., Schwarzschild, M. & Rose, W. K. Infrared spectra of red giant stars. *Astrophys. J.* **140**, 833–852 (1964).

4. Neufeld, D. A., Feuchtgruber, H., Harwit, M. & Melnick, G. J. Infrared Space Observatory observations of far-infrared rotational emission lines of water vapor toward the supergiant star VY Canis Majoris. *Astrophys. J.* **517**, L147–L150 (1999).
5. Millar, T. J., Herbst, E. & Bettens, R. P. A. Large molecules in the envelope surrounding IRC+10216. *Mon. Not. R. Astron. Soc.* **316**, 195–203 (2000).
6. Williams, P. G. & White, G. J. Sub-millimetre molecular lines in the circumstellar envelope of IRC+10216. *Astron. Astrophys.* **266**, 365–376 (1992).
7. Pickett, H. M. *et al.* Submillimeter, millimeter, and microwave spectral line catalog. *J. Quant. Spectrosc. Rad. Transfer* **60**, 883–890 (1998).
8. De Lucia, F. C. & Helminger, P. A. Millimeterwave spectroscopy of active laser plasmas; the excited vibrational states of HCN. *J. Chem. Phys.* **67**, 4262–4267 (1977).
9. Skinner, C. J. *et al.* Modelling the dust and gas outflows from IRC+10216 – I. Ground-based and airborne observations. *Mon. Not. R. Astron. Soc.* **302**, 293–304 (1999).
10. Glassgold, A. E. Circumstellar photochemistry. *Annu. Rev. Astron. Astrophys.* **34**, 241–278 (1996).
11. Willacy, K. & Charnoff, I. Silicon and sulfur chemistry in the inner wind of IRC+10216. *Astron. Astrophys.* **330**, 676–684 (2000).
12. Herpin, F. & Cernicharo, J. O-bearing molecules in carbon-rich proto-planetary objects. *Astrophys. J.* **530**, L129–L132 (2000).
13. Jewitt, D. & Luu, J. X. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 1201–1229 (Univ. Arizona Press, Tucson, 2000).
14. Cernicharo, J. *et al.* The ISO/SWS spectrum of IRC+10216: the vibrational bands of C₂H₂ and HCN. *Astrophys. J.* **526**, L41–L44 (1999).
15. Cernicharo, J. *et al.* The ISO/LWS far infrared spectrum of IRC+10216. *Astron. Astrophys.* **315**, L201–L204 (1996).
16. Ashby, M. L. N. *et al.* An analysis of water line profiles in star formation regions observed by the Submillimeter Wave Astronomy Satellite. *Astrophys. J.* **539**, L115–L118 (2000).
17. Hogerheijde, M. R. & van der Tak, F. F. S. An accelerated Monte Carlo method to solve two-dimensional radiative transfer and molecular excitation. *Astron. Astrophys.* **362**, 697–710 (2000).
18. Melnick, G. J. *et al.* The Submillimeter Wave Astronomy Satellite: science objectives and instrument description. *Astrophys. J.* **539**, L77–L85 (2000).

Acknowledgements

We thank A. Glassgold for help and insights, and T. Millar and E. Herbst for communicating unpublished results from their chemical model of IRC+10216. This work was supported by NASA.

Correspondence and requests for materials should be addressed to G.J.M. (e-mail: gmelnick@cfa.harvard.edu).

Discovery of 12 satellites of Saturn exhibiting orbital clustering

Brett Gladman*, **J. J. Kavelaars†**, **Matthew Holman‡**, **Philip D. Nicholson§**, **Joseph A. Burns§**, **Carl W. Hergenrother||**, **Jean-Marc Petit***, **Brian G. Marsden‡**, **Robert Jacobson¶**, **William Gray#** & **Tommy Grav***

* Observatoire de la Côte d'Azur, B.P. 4229, 06304 Nice Cedex 4, France

† Department of Physics and Astronomy, McMaster University, Hamilton, Ontario L8S 4M1, Canada

‡ Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

§ Department of Astronomy, Cornell University, Ithaca, New York 14853, USA

|| Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA

¶ Jet Propulsion Laboratory, MS 301-150, 4800 Oak Grove Drive, Pasadena, California 91109, USA

Project Pluto, 168 Ridge Road, Bowdoinham, Maine 04008, USA

* Institute of Theoretical Astrophysics, University of Oslo, PB 1029 Blindern, 05315 Oslo, Norway

The giant planets in the Solar System each have two groups of satellites. The regular satellites move along nearly circular orbits in the planet's orbital plane, revolving about it in the same sense as the planet spins. In contrast, the so-called irregular satellites are generally smaller in size and are characterized by large orbits with significant eccentricity, inclination or both. The differences in their characteristics suggest that the regular and irregular satellites formed by different mechanisms: the regular satellites are

believed to have formed in an accretion disk around the planet, like a miniature Solar System, whereas the irregulars are generally thought to be captured planetesimals¹. Here we report the discovery of 12 irregular satellites of Saturn, along with the determinations of their orbits. These orbits, along with the orbits of irregular satellites of Jupiter and Uranus, fall into groups on the basis of their orbital inclinations. We interpret this result as indicating that most of the irregular moons are collisional remnants of larger satellites that were fragmented after capture, rather than being captured independently.

The discovery rate of irregular satellites has greatly increased owing to the availability of large-format charged-coupled device (CCD) arrays mounted on sizeable telescopes and the concomitant capability to handle the enormous data volume produced (30 gigabytes per night). The ability to survey large areas of sky (the irregular satellite systems cover several square degrees around each planet) to depths formerly unattainable with photographic plates has allowed the number of known irregular satellites to nearly quadruple since 1996, increasing from 10 to 39. All five irregulars of Uranus were found in 1997 and 1999 (refs 2 and 3). Saturn's irregular complement was restricted to one, Phoebe, until the discoveries reported here. The number of jovian irregulars, previously eight, has more than doubled. S/1999 J 1 was discovered by Spacewatch and the Minor Planet Center⁴ and recovered (observed again) by B.G. S/2000 J 1 was linked by B.G.M. and G. Williams⁵ to the lost satellite S/1975 J 1 (observed by C. Kowal and E. Roemer) after observations in November 2000 (Sheppard *et al.*), and August 2000 (M.H.) (see ref. 6). In January 2001, Sheppard *et al.*⁶ reported ten more jovian irregulars. Neptune's irregular roster has not grown; we discuss its case below.

We organized an observational campaign to search for new irregulars during Saturn's apparition in late 2000, performing discovery surveys using wide-field CCD mosaics on several telescopes supplemented by recovery observations using moderate-field cameras (Table 1). The discovery searches used mosaic cameras to image the sky surrounding Saturn; three CCD images with about an hour between each image were searched for moons moving across the sky at rates similar to Saturn's. Preliminary observations on the European Southern Observatory (ESO) 2.2-m telescope as Saturn rose in the early morning of 7 August 2000 revealed three satellite candidates. A more thorough search at the Canada–France–Hawaii Telescope (CFHT) on 23 and 24 September recovered these three moons and yielded another dozen candidates. Recovery observations by many workers (Table 1) were then useful in tracking the moons over the next several months. An additional satellite (S/2000 S 11) was discovered on 8 November 2000 outside the initial search area, and confirmed on 23 November. As of February 2001 we have four outstanding clear detections from CFHT that (within the astrometric errors) moved like moons. Our failure to recover three of these is due to their faintness ($m_R \approx 24.2$ – 24.5); the fourth candidate ($m_R \approx 23$) may be a Centaur fortuitously near the planet, because several attempts to recover it at positions consistent with bound orbits have failed.

Throughout this process, independent ephemerides (orbit fitting and positional prediction) provided by B.M., R.J. and W.G. were invaluable in assisting the observing team in satellite recovery. The saturnian irregular satellites have moved over a significant fraction of their orbits since discovery; using small-field cameras at observation intervals of about 1 month meant that fitted ephemerides were essential for retrieval. The currently observed arcs are long enough (as of February 2001) that we expect the calculated orbital elements of the saturnian irregulars (Table 2) to change very little upon recovery after the May 2001 conjunction.

The orbits shown in Fig. 1 were obtained by fitting osculating orbits through published observations and then numerically integrating the trajectories for 1,000–6,000 yr to average out variations produced by solar perturbations. These 'mean elements' (given in

Table 2 for the saturnian irregulars) provide more accurate intrinsic information than osculating elements (for example, for the velocity dispersion between satellite orbits).

Unexpectedly, almost all of the irregulars discovered since 1997 cluster together in easily discernible groupings (Fig. 1). Tracking of the five uranian irregular satellites³ has established that they lie in a single retrograde group, or perhaps two. Saturn exhibits three or four different inclination groupings, one of which includes Phoebe (it is unclear if S/2000 S 8 is part of the Phoebe grouping). Jupiter's eight large irregulars have long been known to fall into prograde and retrograde groups; the new jovian satellites appear to divide the formerly very dispersed retrograde group into at least two subgroups. We argue below that this orbital clumping indicates that most of the current irregulars are collisional fragments rather than independent satellite captures.

Permanent capture from heliocentric onto planetocentric orbits is not easily accomplished because bodies passing near a planet will eventually exit again unless orbital energy can be dissipated or the planet's gravitational reach can suddenly be expanded. Suggested dissipation mechanisms include gas drag as a planetesimal traverses a primitive gas envelope around the planet⁷ or collisions between objects moving within the planet's gravitational grasp⁸. To first order, these mechanisms would produce roughly equal numbers of direct and retrograde satellites. Alternatively, a rapid expansion of the planet's Hill sphere (see below) as the planet's mass grew could trap planetesimals; such a 'pull-down' capture should preferentially yield retrograde bodies^{9,10}. The significant percentage of direct moons seems to imply that some energy-loss mechanism accounts for most, if not all, irregular satellites.

The Hill sphere is the region in which a planet's gravity exceeds the Sun's tidal force, and has radius $R_H = a_p(\mu/3)^{1/3}$, where a_p is the planet's orbital semimajor axis and μ is the planet's mass expressed as a fraction of the Sun's. In numerical simulations^{11–13}, planetocentric retrograde orbits are stable out to about R_H whereas direct ones are only retained out to $0.3–0.5R_H$; more distant retrograde 'quasi-satellites' of Uranus and Neptune may also be stable (see ref. 14). The population of known irregulars does not reach beyond about $0.5R_H$ (Fig. 1), but this may simply indicate that capture was more probable closer in. A heightened capture probability might be due to a greater gas density as a planetesimal penetrates more deeply into the cocoon surrounding a forming giant planet, or perhaps the planetesimal collides with a pre-existing moon (the latter being more populous near the planet). Alternatively, captured but loosely

bound distant irregulars may have been destabilized by passing comets or planetary perturbations over the age of the Solar System. The preference for the more distant jovian satellites to be retrograde, once believed to reflect the greater stability of retrograde orbits, is much less marked in the saturnian clan. We notice that the closer the planet is to the Sun the more distant are its retrograde satellites when expressed in Hill radii (Fig. 1); but when semimajor axes are expressed in planetary radii the retrograde irregulars (excepting Triton) almost all lie within 150–400 radii of their planets. This hints that all protoplanetary envelopes, within which the parent satellites may have been captured, had similar cross-sections in units of planetary radii, supporting gas drag scenarios over pull-down capture because the latter should depend on the tidal Hill radius.

Irregular satellites with orbital inclinations between about 50 and 140 degrees are absent (Fig. 1). This eye-catching feature is related to the Kozai phenomenon^{15,16} in which solar perturbations cause e and I oscillations. As I approaches 90° (either from below, or from above for retrograde objects), large periodic e -oscillations develop. When the peak eccentricity rises high enough, the moon passes through the region occupied by the regular satellites, allowing collisions or gravitational scatterings sufficient to remove the irregular from the system. The resulting 'zone of exclusion' in Fig. 1, calculated for each orbital inclination, is conservative because it assumes initially circular orbits and ignores short-period and resonant (non-secular) perturbations which overlie the smooth Kozai cycle. For example, Jupiter's Pasiphae with $I = 153^\circ$ has e varying between 0.19 and 0.68 due to short-period and resonant perturbations¹⁰ although the Kozai-induced eccentricity amplitude is very small. Thus, the 'zone of exclusion' in Fig. 1 is certainly underestimated.

The tight orbital clustering observed implies that these groups almost certainly arise from the disruption of larger bodies. A test of this break-up model will come if colours within the various irregular groups are homogeneous; for example, Jupiter's irregulars have colours like C and D asteroids, Phoebe is neutral, whereas the large uranian satellites are red^{12,17–19}. The difficult question is whether the disruptions occurred during the capture process itself when the planets formed long ago, or whether intact moons were captured at that time into orbits near the present grouping and these single moons were subsequently shattered and scattered by intruding comets or asteroids during the subsequent ~ 4 Gyr.

For models in which break-up occurs simultaneously with satellite capture, the mean e and I of any cluster are set by the trajectory of the incoming planetesimal and by the effectiveness of

Table 1 Observations of Saturn's irregular satellites in 2000–2001

Date	Telescope	Observers	Search area	Magnitude limit	Satellites observed
7 Aug. 2000	ESO 2.2 m	B.G.	8'–40'	22 R	1*, 2*, 5*
23, 24 Sep. 2000	CFHT 3.6 m	B.G. & J.J.K.	12'–60'	24.5 R	1, 2, 3*, 4*, 5, 6*–10*, 12*
25 Sep. 2000	MDM 2.4 m	L. Allen & T. Rigg	Recovery		3
27 Sep. 2000	Steward 1.5 m	C.W.H. & S. Larson	Recovery		1, 3, 4
29 Sep. 2000	ESO NTT 3.5 m	A. Doressoundiram & J. Romon	Recovery		1, 3
2, 3 Nov. 2000	Steward 2.3 m	C.W.H., R. Whitley & D. Means	Recovery		5, 6
4 Nov. 2000	ESO VLT UT1 8 m	UT1 Science Team†	Recovery		5–9, 12
9 Nov. 2000	Mt Hopkins 1.2 m	M.H.	10'–90'	22 R	11*
23 Nov. 2000	Mt Hopkins 1.2 m	T. Spahr	Recovery		11
24, 25 Nov. 2000	ESO 2.2 m	B.G. & J.-M.P.	†	23 R	10, 11
27 Nov. 2000	Palomar 5 m	J.J.K. & P.D.N.	Recovery		7–10
16–18 Dec. 2000	Kitt Peak 4 m	M.H., B.G. & T.G.	†	23 R	3–6, 10–12
19–23 Jan. 2001	Palomar 5 m	P.D.N. & B.G.	Recovery		3, 4, 6, 8–10, 12
15, 18 Feb. 2001	NOT 2.5 m	T.G. & M.H.	Recovery		2–4, 11

* Indicates the discovery observation.

† Director's discretionary time, measured by B.G.

‡ When combined, these two runs covered a 15° – 120° annulus.

Discovery survey search areas are reported as inner and outer radii of circular annuli (in arcminutes from the planet) with a rough magnitude limit. Scattered light from the planet was a severe impediment during all the mosaic searches resulting in a highly non-uniform limiting magnitude. The CFHT search was complicated by Saturn's proximity to its 'stationary point' where the planet's motion slows as seen from Earth, producing potential confusion between satellites and main-belt asteroids; conversely, this slow motion allowed longer exposures and the sub-arcsecond seeing conditions of those two nights led to excellent astrometry (mean residuals of about $0.1''$) for all detected objects. This allowed us to separate candidate moons from 18 trans-neptunian objects, a Centaur designated 2000 SV₃₃₁, and dozens of main-belt asteroids moving near Saturn's rate. Centaurs constituted our chief source of potential confusion because they could fortuitously be near Saturn's heliocentric distance. We estimate our survey of Saturn's environs to be essentially 'complete' to $m_R = 23$ over the entire Hill sphere, missing $<1\%$ of the search area owing to stellar confusion or gaps in the CCD mosaics. Between $m_R = 23$ and 24.5 only the CFHT survey covered a significant area (≈ 3 square degrees). MDM, Michigan–Dartmouth–MIT; NTT, New Technology Telescope; VLT, Very Large Telescope; and NOT, Nordic Optical Telescope.

gas drag. In this case dispersion about the mean orbit reflects: (1) the variable gas-drag accelerations experienced by particles of different sizes; or (2) fragment velocities at break-up. Gas drag in a circumplanetary nebula could provide a size filter, as large objects will not be slowed enough to allow capture, but the smallest will subsequently rapidly spiral into the planet; sizes between 1–100 km are estimated to be favoured⁷. The hypothesis that gas drag is responsible for both capture and break-up produces contradictions when confronted with the newly available data. Smaller fragments, being more coupled to the gas, should experience faster orbital evolution that spirals them in towards the planet and circularizes their orbits. This is contradicted by the much-larger Phoebe being the closest of its cluster, at Uranus where Caliban and Sycorax (the largest irregular moons) are two of the three closest, at Jupiter where satellites of greatly differing size have the same a , and by the lack of any general correlation between satellite size and e . However, some orbital evolution subsequent to capture seems required to trap at least two of the large jovian irregulars into secular resonances¹⁰.

We thus favour the idea that each cluster is produced by the disruption via cometary bombardment of a single parent object well after the capture. Typical fragment speeds after catastrophic disruption are a few times the speed necessary to escape the parent body's gravity²⁰. Orbital velocities are typically 1 km s^{-1} and the observed e and I dispersions correspond to $\sim 100 \text{ m s}^{-1}$. The largest irregular in most groupings has an escape speed of around 50 m s^{-1} ; the shattered parent was most probably a factor of two larger than this and thus the observed velocity dispersion is compatible with a subsequent break-up. The inner-ring moons of the giant planets are thought to have been broken up by cometary bombardment and accretionally re-assembled several times over the age of the Solar System²¹, but irregular moons will not re-agglomerate because mutual collision times between members of any cluster are longer than the Solar System's age². Timescales for disruption of small bodies in the outer Solar System by cometary bombardment are estimated at about 10^9 – 10^{10} yr (refs 22, 23). Therefore, these moons serve as probes that set upper and lower limits on the cometary bombardment flux, integrated over the lifetime of the Solar System, for otherwise, either no traces of them would have survived until the present day or none of them should yet have been disrupted.

Because our Saturn survey is complete (this prevents undiscovered moons from altering a cumulative size distribution) we can estimate size distributions down to the survey limit at about 3 km. There is less than a factor of two in size between the two largest

bodies in each of the two saturnian direct clusters (Table 2) and inside the Uranus cluster, similar to most asteroid families (thought to be recent break-ups of similarly sized bodies)²⁴. In contrast, Phoebe's group shows a large size gap; Phoebe is around a factor of 15 larger than S/2000 S 1; asteroid families with similar size jumps are modelled as resulting from large-scale cratering events rather than collisional dispersal of the main body. Phoebe's spherical shape²⁵ is, like Vesta, inconsistent with it being a catastrophically disrupted fragment, and their 'families' are more probably crater ejecta; we therefore predict that Cassini spacecraft imaging of Phoebe in 2004 will find a $\sim 50 \text{ km}$ impact structure (below the resolution limit of previous Voyager imaging²⁶).

The richness of the irregular satellite systems is yielding valuable insights into the physical conditions and processes that occurred during the final stages of giant planet formation. When fainter magnitudes are probed with 8–10-m telescopes, or using pencil-beam imaging techniques²⁷, more irregulars will be discovered around all the giant planets, but Neptune is especially interesting, owing to a potentially violent dynamical history related to the capture of Triton²⁸. The largest (about 150 km in diameter) irregulars, like Phoebe or Sycorax, would be of magnitude ~ 22 if moved

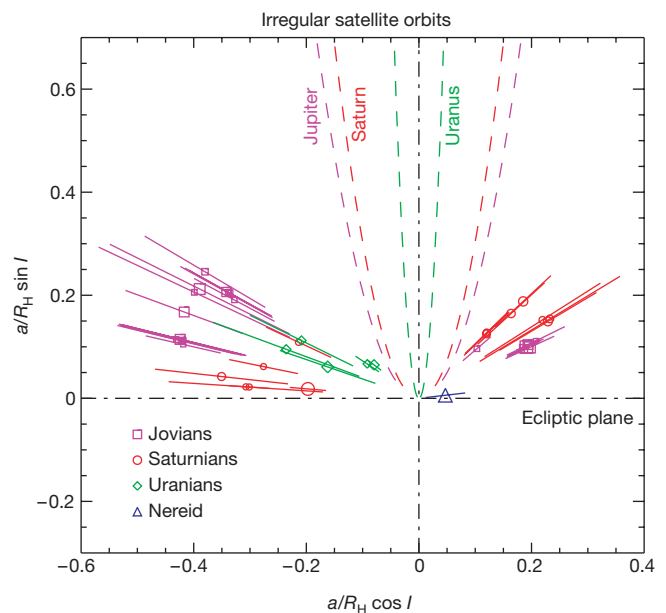


Figure 1 This sketch compares the orbital properties of the irregular satellites of the giant planets. A different colour is used for each planet's satellites. Each orbit's inclination relative to the J2000 ecliptic (approximately the planet's orbital plane about the Sun) is illustrated by the angle from the horizontal; moons in the right quadrant move in the same (direct) sense as their planets orbit the Sun, while those in the left quadrant are retrograde. The radial distance from the origin to the symbol represents the orbital semimajor axis a (given as a fraction of the radius of its planet's Hill sphere R_H). Triton's orbit would appear at the origin at the scale of this diagram. The length of each line illustrates the pericentre-to-apocentre variation in distance that is due to the orbital eccentricity e . The symbol's size is proportional to the logarithm of the satellite's radius, either measured or estimated from its magnitude and an assumed albedo of 6%. Time-averaged orbital elements are used to remove the effect of solar perturbations. The region within which the Kozai mechanism removes nearly polar orbits is indicated by the dashed lines for each planet (see text). Crossing the outermost massive satellites of Jupiter and Saturn (Callisto and Iapetus, respectively) requires $e \approx 0.9$. Typical probabilities for collisions with the outermost regular satellites are about 10^{-6} per orbit; with orbital periods of a few years and the high- e portion of the Kozai modulation lasting a few per cent of the cycle, we estimate that jovian or saturnian irregulars with $I \approx 70$ – 110° would be eliminated in less than the Solar System's age. The uranian loss cone is much narrower because its distant irregular satellites must attain very high e to cross the orbit of its outermost regular satellite Oberon (which is relatively close at only 22 planetary radii).

Table 2 The mean orbital elements of Saturn's irregular satellites

Satellite	a (10^6 km)	e	I ($^\circ$)	P (yr)	m_R	R (km)
34-degree inclination group						
S/2000 S 4	18.2	0.54	33.5	2.53	22.1	7
S/2000 S 10	17.5	0.47	34.7	2.35	23.0	4
S/2000 S 11	15.6	0.55	34.1	2.02	20.5	13
46-degree inclination group						
S/2000 S 2	15.2	0.36	45.1	1.88	21.3	10
S/2000 S 3	17.8	0.27	45.8	2.38	20.1	16
S/2000 S 5	11.3	0.33	46.1	1.23	22.0	7
S/2000 S 6	11.5	0.32	46.6	1.24	22.6	5
S/2000 S 8 alone						
S/2000 S 8	15.7	0.27	153.0	2.00	23.6	3
Phoebe group						
Phoebe	12.94	0.163	174.7	1.51	16.4	110
S/2000 S 1	23.1	0.34	173.1	3.59	21.7	8
S/2000 S 7	20.1	0.45	175.9	2.92	23.9	3
S/2000 S 9	18.5	0.22	167.4	2.57	23.8	3
S/2000 S 12	19.7	0.12	175.8	2.84	23.9	3

Orbital and physical characteristics of the saturnian irregular satellites, grouped by orbital inclination I (referred to the J2000 ecliptic). The mean orbital elements, fitted to a 1,000-yr numerical integration, are listed along with the orbital period P , the apparent magnitude at discovery m_R , and the estimated radius R for an assumed albedo of 0.06 (Phoebe's measured value). Astrometric data can be found in refs 29–35.

to Neptune, whereas the second-largest moons (like Caliban or S/2000 S 3) would have magnitudes of 24–25. Thus if a similar irregular system is present around Neptune, its smaller members were beyond the limits of the deepest known survey³ and Nereid (Neptune's only distant satellite) is only the brightest member of a population waiting patiently to be discovered. □

Received 15 February; accepted 9 May 2001.

1. Peale, S. J. Origin and evolution of the natural satellites. *Annu. Rev. Astron. Astrophys.* **37**, 533–602 (1999).
2. Gladman, B. J. *et al.* Discovery of two distant irregular moons of Uranus. *Nature* **392**, 897–899 (1998).
3. Gladman, B. J. *et al.* The discovery of Uranus XIX, XX, and XXI. *Icarus* **147**, 320–324 (2000); erratum **148**, 320 (2000).
4. Marsden, B. G. S/1999 J 1. *IAU Circ.* **7460** (2000).
5. Marsden, B. G. S/1975 J 1 = S/2000 J 1. *IAU Circ.* **7640** (2000).
6. Green, D. Satellites of Jupiter. *IAU Circ.* **7555** (2001).
7. Pollack, J. B., Burns, J. A. & Tauber, M. E. Gas drag in primordial circumplanetary envelopes: A mechanism for satellite capture. *Icarus* **37**, 587–611 (1979).
8. Colombo, G. & Franklin, F. A. On the formation of the outer satellite groups of Jupiter. *Icarus* **30**, 186–189 (1971).
9. Heppenheimer, T. A. & Porco, C. C. New contributions to the problem of capture. *Icarus* **30**, 385–401 (1977).
10. Saha, P. & Tremaine, S. The orbits of the retrograde jovian satellites. *Icarus* **106**, 549–562 (1993).
11. Hénon, M. Numerical exploration of the restricted problem. VI. Hill's case: Non-periodic orbits. *Astron. Astrophys.* **9**, 24–36 (1970).
12. Hamilton, D. P. & Burns, J. A. Orbital stability zones about asteroids. *Icarus* **92**, 118–131 (1991).
13. Hamilton, D. P. & Krivov, A. V. Dynamics of distant moons of asteroids. *Icarus* **128**, 241–249 (1997).
14. Wiegert, P., Innanen, K. & Mikkola, S. The stability of quasi satellites in the outer solar system. *Astron. J.* **119**, 1978–1984 (2000).
15. Kozai, Y. Secular perturbations of asteroids with high inclination and eccentricity. *Astron. J.* **67**, 591–598 (1962).
16. Kinoshita, H. & Nakai, H. Secular perturbations of fictitious satellites of Uranus. *Celest. Mech.* **52**, 293–303 (1991).
17. Clark, R. N., Fanale, F. P. & Gaffey, M. J. In *Satellites* (eds Burns, J. A. & Matthews, M. S.) 437–492 (Univ. Arizona Press, Tucson, 1986).
18. Luu, J. CCD photometry and spectroscopy of the outer jovian satellites. *Astron. J.* **102**, 1213–1225 (1991).
19. Jarvis, K. S. *et al.* JVI Himalia: New compositional evidence and interpretations for the origin of Jupiter's small satellites. *Icarus* **145**, 445–453 (2000).
20. Farinella, P. *et al.* The injection of asteroid fragments into resonances. *Icarus* **101**, 174–187 (1993).
21. Smith, B. A. *et al.* A new look at the Saturn system. *Science* **215**, 504–537 (1982).
22. Colwell, J. E. & Esposito, L. W. Origins of the rings of Uranus and Neptune. II. Initial distributions of disrupted satellite fragments. *J. Geophys. Res.* **98**, 7387–7401 (1993).
23. Porco, C. C. *et al.* In *Neptune and Triton* (ed. Cruikshank, D. P.) 703–804 (Univ. Arizona Press, Tucson, 1995).
24. Tanga, P. *et al.* On the size distribution of asteroid families: The role of geometry. *Icarus* **141**, 65–78 (1999).
25. Thomas, P. *et al.* Phoebe: Voyager 2 observations. *J. Geophys. Res.* **88**, 8736–8742 (1986).
26. Simonelli, D. *et al.* Phoebe: Albedo map and photometric properties. *Icarus* **138**, 249–258 (1999).
27. Gladman, B. *et al.* Pencil-beam surveys for faint trans-neptunian comets. *Astron. J.* **116**, 2042–2054 (1998).
28. McKinnon, W., Lunine, J. I. & Banfield, D. In *Neptune and Triton* (ed. Cruikshank, D. P.) 807–878 (Univ. Arizona Press, Tucson, 1995).
29. Marsden, B. G. S/2000 S 1 and S/2000 S 2. *IAU Circ.* **7512** (2000).
30. Marsden, B. G. S/2000 S 3 and S/2000 S 4. *IAU Circ.* **7513** (2000).
31. Marsden, B. G. S/2000 S 5 and S/2000 S 6. *IAU Circ.* **7521** (2000).
32. Marsden, B. G. S/2000 S 7, S/2000 S 8 and S/2000 S 9. *IAU Circ.* **7538** (2000).
33. Marsden, B. G. S/2000 S 10. *IAU Circ.* **7539** (2000).
34. Green, D. S/2000 S 11. *IAU Circ.* **7545** (2000).
35. Green, D. S/2000 S 12. *IAU Circ.* **7548** (2000).

Acknowledgements

The Canada–France–Hawaii telescope is operated by the National Research Council of Canada, le Centre National de la Recherche Scientifique de France, and the University of Hawaii. Observations for the ESO 2.2-m telescope and the VLT were collected at the European Southern Observatory, Chile. Observations at the Palomar Observatory were made as part of a continuing collaborative agreement between the California Institute of Technology and Cornell University. Kitt Peak National Observatory, part of the National Optical Astronomy Observatories, is operated by the Association of Universities for Research in Astronomy, Inc. (AURA) under cooperative agreement with the National Science Foundation. The Nordic Optical Telescope is operated on the island of La Palma jointly by Denmark, Finland, Iceland, Norway and Sweden, in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias. We gratefully acknowledge financing from the French Research Ministry ACI Jeunes Chercheurs programme, the Institut National de Science de l'Univers, from the European Southern Observatory, from the NASA Planetary Astronomy programme and the Natural Sciences and Engineering Research Council of Canada.

Correspondence and requests for materials should be addressed to B.G. (e-mail: gladman@obs-nice.fr).

Ion-beam sculpting at nanometre length scales

Jiali Li*, Derek Stein†, Ciaran McMullan‡, Daniel Branton‡, Michael J. Aziz† & Jene A. Golovchenko*†

* Department of Physics, † Division of Engineering and Applied Sciences, and ‡ Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

Manipulating matter at the nanometre scale is important for many electronic, chemical and biological advances^{1–3}, but present solid-state fabrication methods do not reproducibly achieve dimensional control at the nanometre scale. Here we report a means of fashioning matter at these dimensions that uses low-energy ion beams and reveals surprising atomic transport phenomena that occur in a variety of materials and geometries. The method is implemented in a feedback-controlled sputtering system that provides fine control over ion beam exposure and sample temperature. We call the method “ion-beam sculpting”, and apply it to the problem of fabricating a molecular-scale hole, or nanopore, in a thin insulating solid-state membrane. Such pores can serve to localize molecular-scale electrical junctions and switches^{4–6} and function as masks⁷ to create other small-scale structures. Nanopores also function as membrane channels in all living systems, where they serve as extremely sensitive electro-mechanical devices that regulate electric potential, ionic flow, and molecular transport across cellular membranes⁸. We show that ion-beam sculpting can be used to fashion an analogous solid-state device: a robust electronic detector consisting of a single nanopore in a Si₃N₄ membrane, capable of registering single DNA molecules in aqueous solution.

When massive ions with energies of several thousand electronvolts impinge on a surface, an atomic-scale erosion process, called sputtering, removes approximately one atom from the surface for every incident ion^{9–12}. We reasoned that as material is removed from a flat Si₃N₄ surface containing a cavity on its opposite surface (Fig. 1a, top), the flat surface will ultimately intercept the bottom of the bowl shaped cavity, forming a nanopore (Fig. 1a, bottom). Creating a molecular-scale pore requires knowing precisely when to stop the erosion process. The apparatus illustrated in Fig. 1b implements a feedback-controlled ion sputtering system that counts the ions transmitted through the opening pore and extinguishes the erosion process at the appropriate time. The apparatus also controls a number of parameters that we discovered to be important to the ion-beam sculpting process. These include: (1) sample temperature; (2) ion beam duty cycle (defined as the time the beam was on, divided by the sum of the times it was on and off, for pulsed beams); and (3) the instantaneous ion beam flux, F , in ions nm⁻² s⁻¹ when the beam is on the sample.

A sample with a large (~0.1-μm diameter) bowl-shaped cavity was fabricated in a free-standing Si₃N₄ membrane supported on a silicon frame (Fig. 1a). To create a molecular-scale nanopore, the sample was ion-sculpted using 3-keV Ar⁺ ions in the apparatus described above. Surprisingly, experiments on this sample at room temperature did not yield the expected result; a nanopore did not open even after excessively long ion beam exposure. We discovered why by ion-sculpting a membrane that contained a through-hole, rather than a bowl-shaped cavity. At room temperature, the transmitted ion counting rate clearly decreased with increasing ion beam exposure (Fig. 2a), suggesting that the hole was closing rather than opening. The incident beam was switched off when the counting rate fell to 40 counts s⁻¹ (Fig. 2a, inset). Transmission electron microscope (TEM) images of the hole before and after the ion-beam exposure (Fig. 2b and c) revealed that the hole size had indeed

been reduced from 60 nm to 1.8 nm by the growth of a thin membrane (of the order of 10 nm, as deduced from electron microscopy). With sufficient ion beam exposure the nanopore completely closed, and the ion count fell to zero. Thus, in addition to ion-sputter erosion, there must be a lateral atomic flow of matter into the pore, stimulated by the ion beam. That this is a surface, or near-surface, phenomenon is strongly suggested by computer simulations that show the ion beam energy to be deposited within less than 5 nm of the sample surface (J. F. Ziegler and J. P. Biersack, SRIM-2000, available at <http://www.research.ibm.com/ion/ionbeams>).

The flow of matter to the developing nanopore is temperature-dependent (Fig. 3). A transition between pore opening and pore closing was consistently seen at about 5 °C under the ion-beam conditions of Fig. 3. When pore area is plotted versus dose rather than time (Fig. 4), the slope of the data reveals that for continuous beam exposure (grey trace, Fig. 4) the efficiency of pore closing per incident ion is clearly greater at low fluxes than at high fluxes. Figure 4 also shows that a pulsed beam (black data points) closes pores more efficiently than does a continuous ion beam at the same instantaneous flux.

Our demonstrated ability to monitor changing dimensions continuously in the nanometre range while varying experimental parameters provides an unusual opportunity to test microscopic models to account for the observed materials phenomena. Competing processes are probably at work. One is responsible for opening

the pore, probably driven by ion-sputter erosion of the pore edge. This process is probably dominant at low temperatures and high fluxes. Established sputtering phenomenology⁹ could account for this process, although a full understanding will depend on knowing the detailed geometrical shape of the pore, not just its diameter.

Two different views can explain the motion of matter necessary to account for the pore-closing phenomenon. First, a very thin (~5 nm) stressed viscous surface layer may be created by the energy and matter deposited by the ion beam. An enhanced collective motion, driven by a reduced viscosity and/or enhanced stress owing to implantation effects or surface tension, causes the layer to relax. A similar model has been invoked for atomic transport in sputter-induced rippling (refs 13–15).

Second, we can account for our observations with a model in which incident ions both create and annihilate excess, independent and mobile surface 'adatoms' (for example, atoms or molecular clusters) that can diffuse to the pore. Adatom diffusion has been successfully invoked in modelling sputter-induced ripple formation on Si(001)^{16,17}.

We propose that the concentration of surface adatoms $C(\mathbf{r}, t)$, is governed by a two-dimensional diffusion equation:

$$\frac{\partial}{\partial t} C(\mathbf{r}, t) = F Y_a - F C \sigma - \frac{C}{\tau_{\text{trap}}} + D \nabla^2 C \quad (1)$$

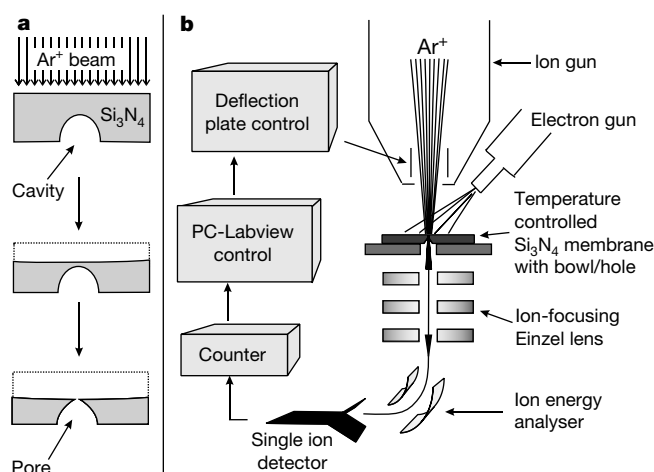


Figure 1 Strategy to make nanopores using argon ion-beam sputtering. **a**, Sputtering removes material from a free-standing Si_3N_4 membrane with a cavity. **b**, Feedback-controlled ion-beam sculpting apparatus housed in a high-vacuum chamber. **a**, A 500-nm-thick low-stress (~200 MPa tensile) Si_3N_4 film was deposited on a (100) silicon substrate by low-pressure chemical vapour deposition²⁵. Photolithography and directional wet chemical etching of silicon were used to create a free-standing $25 \mu\text{m} \times 25 \mu\text{m}$ Si_3N_4 membrane²⁶. Either a bowl-shaped cavity (**a**), or a single initial pore of ~0.1 μm diameter (not shown), was created near the centre of the membrane using, respectively, reactive ion etching²⁷ or a focused ion beam (FIB) machine²⁶. **b**, A differentially pumped ion gun (VG Microtech model EX05) exposes the sample surface to an Ar^+ beam, ~0.2 mm in diameter. A Channeltron (Gallileo Optics) electron-multiplier-style single-ion detector, positioned after the sample, counts transmitted ions. Deflection plates at the exit port of the ion gun could deflect the beam off the sample or pulse the ion beam on and off the sample. A focusing Einzel lens and 60° electrostatic deflection system between sample and detector are used to suppress electron, ion and X-ray backgrounds. A 50-eV electron gun (Kimball Physics Model FRA-2x1-1) floods the sample to neutralize surface charging. A liquid-nitrogen-cooled shroud surrounds the sample and Einzel lens and a quadrupole mass spectrometer, connected to the 10^{-9} torr turbo-pumped vacuum chamber, monitors residual gas composition. A thermocouple monitors the sample holder temperature, which is adjusted with cold nitrogen gas and a resistance heater.

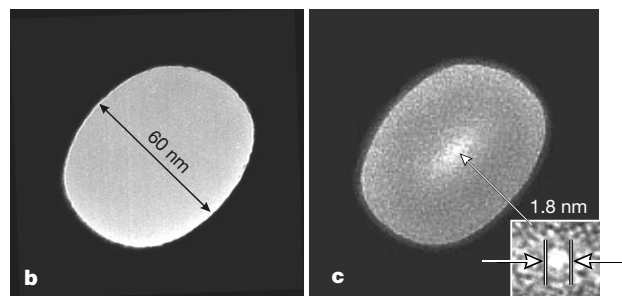
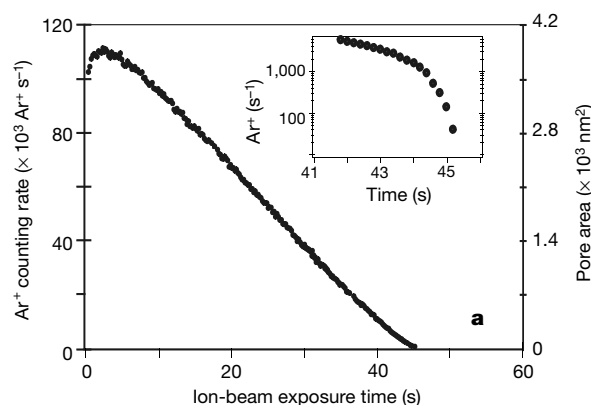


Figure 2 Sculpting a nanopore. **a**, Transmitted ion count rate (left axis) and pore area (right axis) versus integrated time the ion beam is on the 28 °C sample. **b**, TEM image of initial 61-nm diameter pore made by FIB in a 500-nm Si_3N_4 membrane. **c**, TEM image of the same sample after Ar^+ ion-beam exposure. Energy-dispersive analysis of X-rays in the TEM reveals the presence of Si and N in the membrane that has filled the pore, although the precise composition has not been quantified. Because the transmitted ion current is directly proportional to the area of the pore, the instantaneous pore area indicated in all figures was calculated by multiplying the initial pore area (determined by TEM) by the ratio of the instantaneous to initial transmitted ion current. Temperature, 28 °C. Flux, $28 \text{ Ar}^+ \text{ s}^{-1} \text{ nm}^{-2}$. Duty cycle, 200 ms/1 s.

where r and t are surface position and time. D is the adatom surface-diffusion coefficient and F the incident ion-beam flux. $\partial C/\partial t$ depends on: (1) a generation rate, taken proportional to the incoming flux through Y_a ; (2) an annihilation rate proportional to F and C through an annihilation cross-section σ , and an annihilation rate due to trapping at surface defects, assumed to be proportional to C through τ_{trap}^{-1} with τ_{trap} a lifetime. Term (2) causes the model to exhibit a reduced efficiency of pore closing per incident ion at increased fluxes.

The pore boundary is taken to be a perfect sink for adatoms, which are there transformed to a thin layer of accumulating matter that accounts for pore closure. (A vacancy source at the pore, instead, could also produce this effect.) Our boundary condition could arise from capillary forces, which under equilibrium conditions drive the pore to close when the radius of curvature in the plane of the membrane is smaller than that in an axial section, and vice versa. The pore size at which the actual transition between opening and closing occurs could be altered by the influence of ion bombardment on the energetics or kinetics governing atomic transport.

From steady-state solutions to equation (1) the diffusional flux into the pore can be obtained. There results a characteristic distance from the pore edge, X_m , within which adatoms are more likely to reach the pore than be annihilated by traps or ion erosion, where:

$$\frac{1}{X_m^2} = \frac{1}{D\tau_{\text{trap}}} + \frac{\sigma}{D}F \quad (2)$$

Adatoms beyond X_m are more likely to be annihilated before they reach the pore.

In low-temperature experiments, where diffusional contributions to pore closure are presumably frozen out (see Fig. 3), the sputter-erosion process that contributes to pore opening can be determined. We include this process (assumed to be temperature-independent) taking $Y_a \approx 1$ and a pore thickness of about 10 nm, and fit X_m for each incident flux at 28 °C (where pores close); this yields the solid curves in Fig. 4. As expected, X_m increases with decreasing flux, which accounts for the enhanced pore-closing efficiency at low flux. The resulting X_m values are consistent with equation (2), as shown in the inset of Fig. 4, from which $D \approx 10^3 \text{ nm}^2 \text{ s}^{-1}$ is extracted from a linear fit using $\sigma \approx 0.1 \text{ nm}^2$ as a reasonable estimate.

The model qualitatively explains the pulsed ion-beam observations. When the ion beam is off, adatoms remain on the surface, but the adatom annihilation channel associated with the incident beam flux disappears. Thus, after the beam is extinguished, the remaining adatoms may diffuse to the pore periphery from greatly increased X_m . This can significantly increase the efficiency per ion for pore closing. The enhanced pore closing with increasing temperature

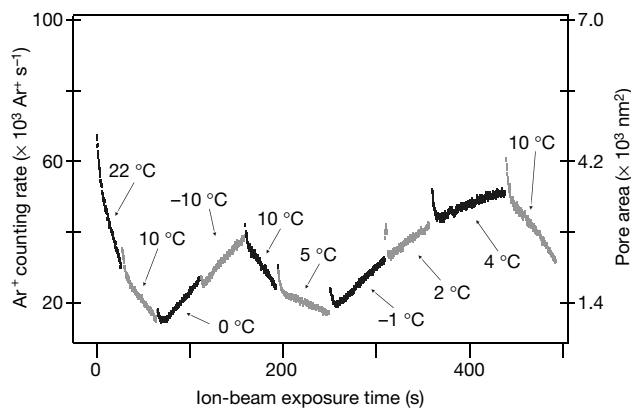


Figure 3 Temperature dependence of ion-beam sculpting. Successive data sets at different temperatures (shown) are delimited by their alternate black and grey coloration. Flux, $14 \text{ Ar}^+ \text{ s}^{-1} \text{ nm}^{-2}$. Duty cycle, 200 ms/1 s.

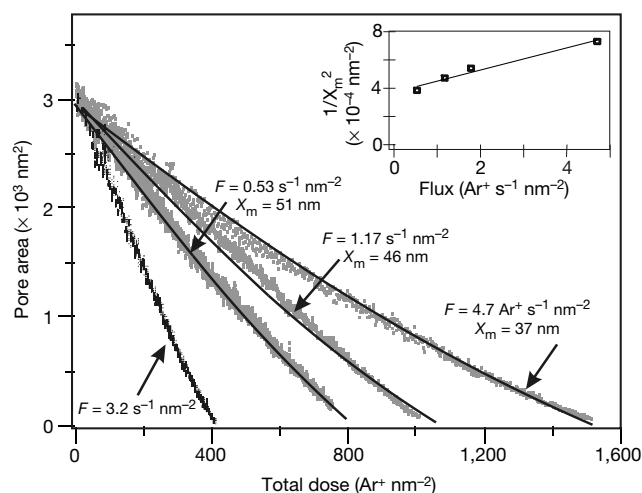


Figure 4 Flux dependence of ion-beam sculpting. Pore area versus total dose for samples exposed at different instantaneous fluxes, F , to a continuous beam (grey traces), or a pulsed beam (black traces). Duty cycle, 100 ms/1 s. The plotted black curves overlying the grey data points are predicted from the diffusion model under steady-state conditions (see text). The inset plots $1/X_m^2$ versus flux, from which D is extracted. Temperature, 28 °C.

can be accounted for by a thermally activated adatom diffusion coefficient.

The diffusion model presented above is phenomenological and contains idealizations and assumptions connected with our ignorance of many microscopic properties of matter under ion beam exposure. Nevertheless, extensions of the studies demonstrated here, using pulsed as well as continuous beam exposures at different temperatures, should permit the determination of materials-specific parameters like D , τ_{trap} , Y_a and σ that will enable fabrication of useful nanoscale devices.

To demonstrate such a device we sculpted a nanopore in a Si_3N_4 membrane for use as a single-molecule electronic detector of DNA. Proteinaceous nanopores, or channels, have been inserted into lipid bilayers in aqueous solutions where they serve as electronic sensors to identify and characterize single molecules^{18–20}. But proteins in lipid bilayers are labile and the channel diameters they provide cannot easily be adjusted. Robust, solid-state nanopores, fashioned to any desired diameter, could yield new data and understanding of transport in confined spaces, and will make it possible to create robust single-molecule-sensing devices to characterize molecules of

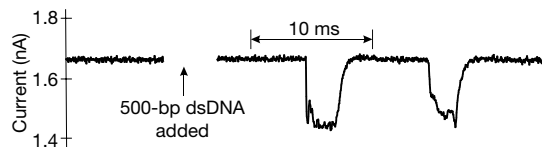


Figure 5 Molecular events in a nanopore detector. A Si_3N_4 membrane with a 5-nm pore separated two compartments filled with saline solution (1 M KCl, 10 mM Tris-HCl, 1 mM EDTA, pH 8.0). Initially, with only the saline solution in the compartments, a 120-mV bias between AgCl electrodes in each compartment resulted in a constant ionic current of 1.66 nA through the nanopore. This was consistent with the known conductivity of the ionic solution, assuming a pore length of around 10 nm. After adding double-stranded DNA, 500 base pairs long, to the negatively biased compartment, and allowing time for diffusion, intermittent current blockades (two are illustrated) were observed. Si_3N_4 membranes with holes of about 100 nm in diameter that were completely closed by ion-beam sculpting produced 20 GΩ seals.

DNA and other biopolymers at unprecedented speeds. Using electrophysiology techniques^{8,21}, we tested one of our robust, electrically quiet, 5-nm-diameter pores with double-stranded DNA. After applying a voltage bias that would draw the negatively charged DNA molecules through the nanopore, we observed diminutions of the ionic current (Fig. 5), reminiscent of the ionic-current blockages observed when single strands of DNA are translocated through the channel formed by α -haemolysin in a lipid bilayer^{22,23}. Because no such blockages were seen during one hour of monitoring before adding DNA, and because the blockages ceased when the voltage bias was reversed, we attribute these blockages to interactions of individual DNA molecules with the nanopore. The duration of these blockades was on the order of milliseconds, and they consistently exhibited current reductions to 88% of the open-pore value. This last value is commensurate with translocation of a rod-like molecule whose cross-sectional area is 3–4 nm² (ref. 24).

The experimental observations, model considerations and the demonstration of an electronic device show that ion-beam sculpting represents a promising new approach to nanoscale fabrication. With feedback control, reproducibility does not depend on precisely matching all conditions and starting dimensions. The method should be useful for fabricating a variety of nanoscale semiconductor devices, as similar sculpting phenomena have been observed for geometries such as thin slits, trenches and crosses, in several materials, like SiO₂, Si and Al. Furthermore, next-generation ion-source arrays and mask technologies (see <http://www-afrd.lbl.gov/ibt.html>) combined with multichannel ion detectors will allow highly parallel applications of nanoscale ion sculpting methods. □

Received 6 February; accepted 11 May 2001.

- Ito, T. & Okazaki, S. Pushing the limits of lithography. *Nature* **406**, 1027–1031 (2000).
- Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408**, 541–548 (2000).
- Stupp, S. I. & Braun, P. V. Molecular manipulation of microstructures: Biomaterials, ceramics, and semiconductors. *Science* **277**, 1242–1248 (1997).
- Reed, M. A., Zhou, C., Deshpande, M. R. & Muller, C. J. The electrical measurement of molecular junctions. *Ann. NY Acad. Sci.* **852**, 133–144 (1998).
- Zhou, C., Deshpande, M. R. & Reed, M. A. Nanoscale metal/self-assembled monolayer/metal heterostructures. *Appl. Phys. Lett.* **71**, 611–613 (1997).
- Ralph, D. C., Black, C. T. & Tinkham, M. Spectroscopic measurements of discrete electronic states in single metal particles. *Phys. Rev. Lett.* **74**, 3241–3244 (1995).
- Deshmukh, M. M., Ralph, D. C., Thomas, M. & Silcox, J. Nanofabrication using a stencil mask. *Appl. Phys. Lett.* **75**, 1631–1633 (1999).
- Hille, B. *Ionic Channels and Excitable Membranes* (Sinauer, Sunderland, Massachusetts, 1992).
- Johnson, R. E. & Shou, J. Sputtering of inorganic insulators. *K. Danske Vidensk. Selsk. Mat.-fys. Meddr.* **43**, 403–494 (1993).
- Sigmund, P. Introduction to sputtering. *K. Danske Vidensk. Selsk. Mat.-fys. Meddr.* **43**, 7–26 (1993).
- Nenandovic, T., Perrillon, B., Bogdanov, Z., Djordjevic, Z. & Milic, M. Sputtering and surface topography of oxides. *Nucl. Instrum. Methods B* **48**, 538–543 (1990).
- Gnaser, H. *Ion Irradiation of Solid Surfaces* (Springer, Berlin/Heidelberg, New York, 1999).
- Mayer, T. M., Chason, E. & Howard, A. J. Roughening instability and ion-induced viscous relaxation of SiO₂ surfaces. *J. Appl. Phys.* **76**, 1633–1643 (1994).
- Carter, G. Viscoelastic buckling and plastic-flow deterministic mechanisms for ripple initiation on ion bombarded amorphous solids. *Surf. Interf. Anal.* **25**, 952–954 (1997).
- Brongersma, M. L., Snoeks, E., Dillen, T. V. & Polman, A. Origin of MeV ion irradiation-induced stress changes in SiO₂. *J. Appl. Phys.* **88**, 59–64 (2000).
- Erlebacher, J., Aziz, M. J., Chason, E., Sinclair, M. B. & Floro, J. A. Spontaneous pattern formation on ion bombarded Si(001). *Phys. Rev. Lett.* **82**, 2330–2333 (1999).
- Erlebacher, J., Aziz, M. J., Chason, E. & Aziz, M. J. Nonlinear amplitude evolution during spontaneous patterning of ion-bombarded Si(001). *J. Vacuum Sci. Technol. A* **18**, 115–120 (2000).
- Kasianowicz, J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl Acad. Sci. USA* **93**, 13770–13773 (1996).
- Bezrukov, S. M., Vodyanoy, I. & Parsegian, V. A. Counting polymers moving through a single ion channel. *Nature* **370**, 279–281 (1994).
- Gu, L. Q., Braha, O., Conlan, S., Cheley, S. & Bayley, H. Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adaptor. *Nature* **398**, 686–690 (1999).
- Hammill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Eur. J. Physiol.* **391**, 85–100 (1981).
- Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E. & Deamer, D. W. Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys. J.* **77**, 3227–3233 (1999).

- Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl Acad. Sci. USA* **97**, 1079–1084 (2000).
- Bezrukov, S. M. Ion channels as molecular coulter counters to probe metabolite transport. *J. Membr. Biol.* **174**, 1–13 (2000).
- Wolf, S. & Tauber, R. N. *Silicon Processing for the VLSI Era* 149–224 (Lattice, Sunset Beach, California, 2000).
- Rai-Choudhury, P. (ed.) *Handbook of Microlithography, Micromachining, and Microfabrication* 41–97, 153–195 (SPIE-International Society for Optical Engineering, Bellingham, Washington, 1997).
- Ralls, K. S., Buhrman, R. A. & Tiberio, R. C. Fabrication of thin-film metal nanobridges. *Appl. Phys. Lett.* **55**, 2459–2461 (1989).

Acknowledgements

This work was supported by the US Defense Advanced Research Projects Agency, the National Science Foundation and the US Department of Energy.

Correspondence and requests for materials should be addressed to J.A.G. (e-mail: govchenko@physics.harvard.edu).

Ordered nanoporous arrays of carbon supporting high dispersions of platinum nanoparticles

Sang Hoon Joo*, Seong Jae Choi*, Ilwhan Oh†, Juhyouon Kwak‡, Zheng Liu‡, Osamu Terasaki‡§ & Ryong Ryoo*

* National Creative Research Initiative Center for Functional Nanomaterials and Department of Chemistry, Korea Advanced Institute of Science and Technology, Taejeon 305-701, Korea

† Electrochemistry Laboratory, Department of Chemistry, Korea Advanced Institute of Science and Technology, Taejeon 305-701, Korea

‡ CREST, Japan Science and Technology Corporation, Department of Physics, Tohoku University, Sendai 980-8578, Japan

§ Department of Physics and Center for Interdisciplinary Research, Tohoku University, Sendai 980-8578, Japan

Nanostructured carbon materials are potentially of great technological interest for the development of electronic^{1,2}, catalytic^{3,4} and hydrogen-storage systems^{5,6}. Here we describe a general strategy for the synthesis of highly ordered, rigid arrays of nanoporous carbon having uniform but tunable diameters (typically 6 nanometres inside and 9 nanometres outside). These structures are formed by using ordered mesoporous silicas as templates, the removal of which leaves a partially ordered graphitic framework. The resulting material supports a high dispersion of platinum nanoparticles, exceeding that of other common microporous carbon materials (such as carbon black, charcoal and activated carbon fibres). The platinum cluster diameter can be controlled to below 3 nanometres, and the high dispersion of these metal clusters gives rise to promising electrocatalytic activity for oxygen reduction, which could prove to be practically relevant for fuel-cell technologies. These nanomaterials can also be prepared in the form of free-standing films by using ordered silica films as the templates.

Various production methods⁷ such as arc discharge, laser ablation, chemical vapour deposition, and template synthesis techniques⁸ are used to obtain carbon nanotubes in the single-wall, multi-wall or disordered-wall form. In general, during synthesis of the nanotubes, the tube diameters are very difficult to control. The carbon nanotubes are obtained as a powder, with separate or entangled nanotubes that exhibit a broad distribution in tube diameters. Some of the single-wall nanotubes undergo self-organization to a bundle⁹. However, the organization is achieved through weak van der Waals interactions, so that the bundle cannot be considered as a system with rigid structural periodicity. Here we

show that a periodic array of uniform ordered nanoporous carbon can easily be synthesized with tunable pore diameters and rigid structural order (Fig. 1), using the mesoporous aluminosilicate molecular sieves known as SBA-15¹⁰ as templates.

The mesoporous SBA-15 can be produced in a large quantity by the cooperative assembly process which takes place between silica species and the amphiphilic poly(alkylene oxide)-type triblock copolymers in aqueous or organic medium¹⁰. The SBA-15 silica is constructed with a hexagonal array of nanotubes with uniform diameter. The diameter of the silica tubes can at present be controlled over the range of 6–15 nm depending on synthesis conditions¹¹. These tubes are randomly interconnected through complementary pores less than 3.5 nm in diameter. Typically, in our experiments, the SBA-15 silica was synthesized in powder form (particle size of $\sim 1\ \mu\text{m}$), using tetraethoxysilicon [$\text{Si}(\text{C}_2\text{H}_5\text{O})_4$] and the $\text{HO}(\text{C}_2\text{H}_4\text{O})_{20}(\text{C}_3\text{H}_6\text{O})_{70}(\text{C}_2\text{H}_4\text{O})_{20}\text{H}$ triblock copolymer (ref. 10). The resulting SBA-15 silica had cylindrical channels 9 nm in diameter. The SBA-15 silica was converted to an aluminosilicate form with an Si/Al ratio of 20, following the postsynthesis incorporation procedure¹². The pore volume of the aluminosilicate SBA-15 was filled with furfuryl alcohol ($\text{C}_5\text{H}_6\text{O}_2$) by the incipient-wetness technique. The aluminosilicate SBA-15 containing the furfuryl alcohol was typically heated for 3 h at 80 °C, which resulted in the aluminosilicate acid-catalytic polymerization of the furfuryl alcohol in the form of a layer coated on the pore walls. The remaining furfuryl alcohol without polymerization in the core of the template pores was removed by subsequent evacuation at 80 °C. The polymerized furfuryl alcohol was converted to carbon inside the SBA-15 template by pyrolysis under vacuum while increasing temperatures up to 1,100 °C. The template was then removed with hydrofluoric acid or aqueous NaOH solution. Elemental analysis and energy dispersive X-ray analysis indicated that a negligible amount of silica remained.

The structure of the synthesized carbon is composed of ordered nanoporous carbon, which was originally formed inside the cylindrical nanotubes of the SBA-15 template. Even once the template has been completely removed, the ordered nanoporous carbon is rigidly interconnected into a highly ordered hexagonal array by carbon spacers, which are formed inside the complementary pores between adjacent cylinders. The mesoscopic structural order between the carbon cylinders gives rise to the appearance of more than five Bragg X-ray diffraction (XRD) lines at small scattering

angles below $2\theta = 5^\circ$ (Fig. 2a). However, the carbon atoms in the frameworks do not have sufficiently long-range atomic order to exhibit XRD peaks at larger scattering angles. Owing to the short-range atomic order, as in partially ordered graphite, ring patterns and fragmented graphitic lattice fringes appear in the electron diffraction pattern and high-resolution transmission electron micrographs (Fig. 2b). We note that whereas our previous work using sucrose as a carbon source resulted in the formation of rod-type carbons^{13,14}, the polymerization of furfuryl alcohol by the designed catalytic function of the aluminosilicate frameworks leads to the formation of these pipe-like carbons. The pore-size distribution curve (obtained by the N_2 adsorption isotherm following the Barrett–Joyner–Halenda method with calibration¹⁵) exhibited two sharp peaks, with the maxima corresponding to the inside diameter of the carbon cylinders (5.9 nm in Fig. 5 of the Supplementary Information) and the pores formed between the adjacent cylinders (4.2 nm), respectively. Here, the outside diameter of the cylinders is controllable by the choice of a template SBA-15 with suitable pore diameter. But the inside diameter, and consequently the wall thickness, is controllable in several ways. One way to do this is by changing the amount of the polymerized furfuryl alcohol using different polymerization temperatures and reaction times. Another way is to add more furfuryl alcohol after the initial polymerization that results in the loss of water.

In such a nanostructured carbon, the pores or channels behave as individual nanoscale reactors so that chemical reactions are confined to take place inside the pores, with only limited diffusion between them. This phenomenon clearly suggests a method of producing nanoscale materials¹⁶. It also means that nanostructured carbon can be used to support small metal clusters for catalytic¹⁷ and electrocatalytic applications¹⁸. High metal dispersion (that is, the large fraction of atoms located at the surface of the small metal clusters) is an important design factor for catalysts. High metal dispersion is useful, not only because it saves expensive metal, but also in controlling structure sensitivity (for example, the catalytic selectivity can be changed by decreasing the cluster size). Platinum clusters as small as 1 nm in diameter can be prepared encased in faujasite-type zeolites, and these clusters exhibit high stability up to 400 °C¹⁹. However, until now, such high metal dispersion combined with stability has been very difficult to obtain with carbons. There are well known methods for producing high Pt dispersions on carbon, but the resulting materials are not used as catalysts because

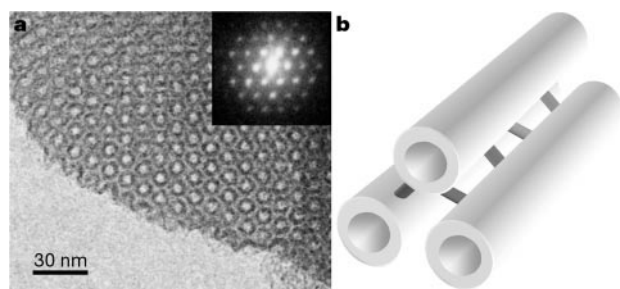


Figure 1 Ordered nanoporous carbon obtained by template synthesis using ordered mesoporous silica SBA-15. **a**, TEM image viewed along the direction of the ordered nanoporous carbon and the corresponding Fourier diffractogram. The image was taken by a JEOL JEM-4000EX instrument (operated at 400 kV). **b**, Schematic model for the carbon structure. The structural model is provided to indicate that the carbon nanopores are rigidly interconnected into a highly ordered hexagonal array by carbon spacers. The outside diameter of the carbon structures is controllable by the choice of a template SBA-15 aluminosilicate with suitable diameter; the inside diameter is controllable by the amount of the carbon source.

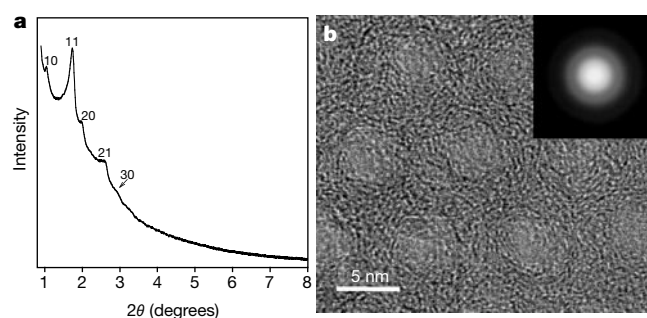


Figure 2 Long-range and short-range order in the structure of ordered nanoporous carbon material. **a**, X-ray diffraction pattern indicating the hexagonal order between carbon cylinders. The XRD pattern was taken by a Rigaku D/MAX-III instrument (operated at 3 kW). The (10) diffraction peak is lower than (11) in intensity, owing to the diffraction interference between the walls and the spacers interconnecting adjacent cylinders. **b**, High-resolution TEM image of carbon cylinders and the corresponding Fourier diffractogram. The image was taken by a JEOL JEM-4000EX instrument (operated at 400 kV). The partially ordered graphitic wall structure not only causes the graphitic lattice fringes to appear in the high-resolution TEM image but also the electron-diffraction ring pattern.

they tend to sinter through Ostwald ripening processes fairly quickly. This difficulty motivated numerous studies to improve metal dispersions on carbons, mainly through optimization of the metal-supporting procedures or functionalization of the carbon surface¹⁷.

However, we found that such small Pt clusters can be prepared very easily by using nanostructured carbon as a support, after a simple incipient-wetness procedure without any surface functionalization. The supported Pt clusters were prepared with an acetone solution of hexachloroplatinic acid (see Methods) and characterized with extended X-ray absorption fine structure (EXAFS), high-resolution TEM and hydrogen chemisorption. The EXAFS result (a Pt–Pt coordination number of 5.4 as shown in Fig. 6 of the Supplementary Information) and the hydrogen chemisorption data (1.5 H atoms per Pt atom) indicate that, when the Pt loading is 2 wt%, the Pt clusters are even smaller than the extremely well dispersed Pt clusters encased in the supercage (1.3 nm in diameter) of NaY zeolites¹⁹. The cluster size increases with the metal loading, but the extent of the increase is much smaller than for conventional porous carbons. Even when the Pt loading is increased to the same weight of carbon (that is, 50 wt% Pt of the total weight), the Pt clusters show a very narrow particle-size distribution around 2.5 nm (Fig. 3). On the other hand, in the case of other porous carbons such as carbon black, activated charcoal and activated carbon fibre, the same experiments have resulted in the formation of much larger Pt particles with a wide distribution of diameters ranging up to 30 nm.

Nanostructured carbon, capable of supporting Pt clusters with such high and uniform dispersion, may find the most suitable application in fuel cell systems²⁰ and as hydrophobic catalysts for bulky organic compounds¹⁷. In fuel cells especially, the high loading of expensive Pt on carbon black has severely limited their widespread use. With this motivation, we compared the electrocatalytic activity of carbon-supported Pt in O₂ reduction with the activities of other carbons, using a rotating-disk electrode. The electrode was

prepared by coating 1.5 µg of catalyst (Pt plus carbon) as a thin film onto glassy carbon (see Methods). The electrocatalytic mass activities, that is, the electrocatalytic currents per gram of Pt, were measured at 900 mV with respect to the normal hydrogen electrode, where the reaction can be considered to occur within the kinetic controlled regime^{21–23}. The mass activities thus obtained with the 20–50 wt% Pt loadings are much higher than those of the Pt/carbon black samples onto which the Pt particles were supported, using the same procedure as for the ordered nanoporous carbon in the present work (see Fig. 4a). The activity-versus-loading plot for the Pt/nanoporous carbon shows a surprisingly high peak activity amounting to 100 A per g Pt at the 33-wt% Pt loading. Such a high activity indicates that several times higher electrocatalytic currents can be generated using the same amount of Pt in the case of the ordered nanoporous carbon. Other investigations have reported a peak activity ranging from 1.4 to 25 A per g Pt at 900 mV under similar conditions^{21–23}. Furthermore, the catalytic current of the Pt ordered nanoporous carbon electrode began to rise much more sharply at a more positive potential, which directly improved the cell efficiency (Fig. 4b). The advantage seemed to come from the uniformity of, and the decrease in, the Pt cluster size when Pt is supported on the nanostructured carbon. Nanostructured carbon with such properties may also be useful for the construction of fuel cell anodes that operate with direct methanol oxidation, where high Pt loading is essential²⁴.

In addition to the powder form, it is also possible to prepare nanostructured carbon in the form of free-standing thin films by using porous silica films as templates. The carbon films thus synthesized retain the shape of the template films (see Fig. 7 in the Supplementary Information), as well as the hexagonally ordered arrangement of the ordered nanoporous carbon in the same way as in the powder form. The development of nanoporous carbons of such ordered film types may be of great importance in separation technologies using membranes, whereas chemical vapour deposition techniques to prepare carbon films largely fail to produce such highly ordered nanotubes or nanopores. The synthesis technique for carbon may be combined with the recently developed sol–gel methods that are suitable for processing the ordered mesoporous silica templates into the forms of oriented fibres²⁵ and variously micro-patterned thin films^{26,27}. It should also be possible to fabricate micro-bundles of the parallel ordered nanoporous carbon if the mesoporous silica templates can be patterned by micro ink-jet printing²⁸ of a surfactant containing a silica sol onto the surface of substrates. □

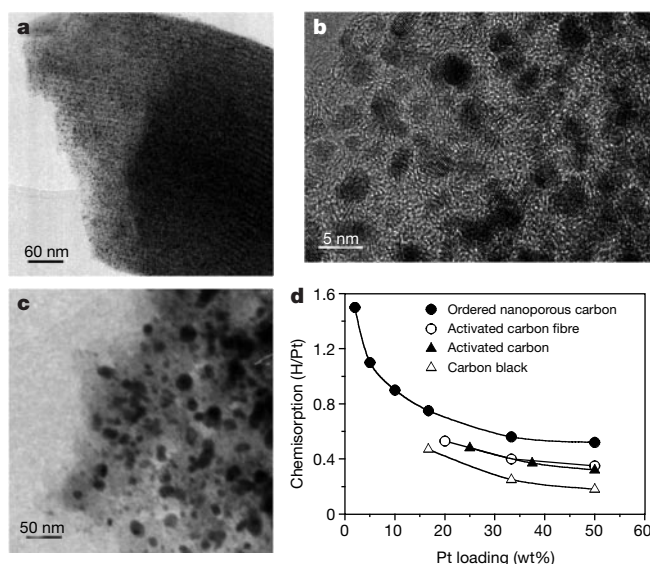


Figure 3 Transmission electron microscope images of the carbon samples supporting the same amount of platinum as the carbon weight. **a, b**, Ordered nanoporous carbon with specific surface area 2,000 m² g^{−1} as measured by the Brunauer–Emmett–Teller (BET) method; **c**, carbon black (Vulcan XC-72). **d**, Hydrogen chemisorption data (that is, the average number of H atoms chemisorbed per Pt atom) for various kinds of carbons supporting Pt: ordered nanoporous carbon (filled circles), carbon black (empty triangles), activated carbon (Darco KB, BET area = 1,500 m² g^{−1}) (filled triangle) and activated carbon fibre (Osaka ACF A-15, BET area = 1,500 m² g^{−1}) (empty circles).

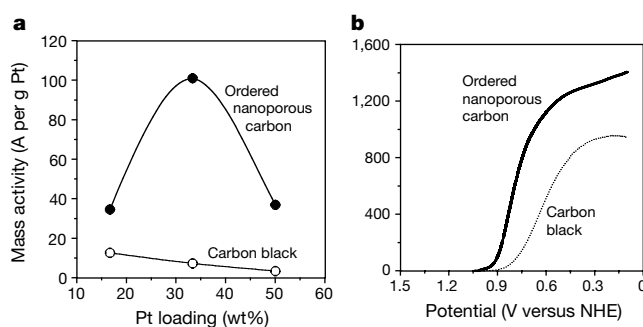


Figure 4 Electrocatalytic mass activities of Pt/carbon catalysts for the O₂ reduction. **a**, Catalytic activity in amperes per gram of Pt measured at a potential of +0.900 V with respect to the normal hydrogen electrode and at a rotating speed of 10,000 r.p.m. in 0.1 M HClO₄ saturated with O₂. **b**, The activity–potential relation for 33-wt% Pt/carbon, obtained at 10,000 r.p.m. with a scan rate of 50 mV s^{−1}. The activities were measured on a rotating disk electrode coated with Pt/carbon. The activity of the 33-wt% Pt ordered nanoporous carbon at +0.900 V is 100 ± 16 A per Pt at the 95% confidence level.

Methods

Preparation of Pt clusters, hydrogen chemisorption and EXAFS

One gram of carbon was impregnated with 2 ml acetone containing H_2PtCl_6 , drop by drop with vigorous agitation. The amount of H_2PtCl_6 in the solution was varied, depending on the desired metal loading. After being dried in a 60 °C oven, the impregnated carbon sample was heated in a H_2 flow while increasing the temperature from room temperature to 300 °C over 2 h. The sample was subsequently outgassed for 2 h at 300 °C, for the desorption of H_2 from the resultant Pt clusters. Hydrogen adsorption isotherms were measured at room temperature, *in situ* on the Pt clusters, using a volumetric adsorption apparatus. The hydrogen chemisorption (the number of H atoms per Pt atom) was determined by the extrapolation of the adsorption isotherm in the range of 10–30 kPa to zero pressure. For EXAFS, the sample that was outgassed at 300 °C was cooled to room temperature and exposed to air. About 0.1 g of the powder sample was pressed into a disk 10 mm in diameter, using polyethylene powder as a binder, and subsequently treated with H_2 at 80 °C. The EXAFS was measured at the Pt L_{III} edge at room temperature under H_2 atmosphere²⁶, using the BL 10B facility at the Photon Factory in Tsukuba. Analysis of the EXAFS data was carried out by standard methods using the UWXAFS2 program package as in ref. 19.

Preparation of electrodes and electrocatalytic activity measurements

Twenty milligrams of Pt/C powder and 0.40 ml ethanol containing 5.0 wt% Nafion were ultrasonically dispersed in 100 ml distilled water. A 30- μl portion of the resultant ink was dropped onto an electrode surface, which was composed of a glassy carbon core, 3 mm in diameter; the surrounding insulation area was 6 mm in diameter. The ink was carefully dried in a 70 °C oven so that Pt catalysts could be uniformly coated over the entire cross-section of the 6-mm diameter area. The electrocatalytic current was measured at room temperature and a rotating speed of 10,000 r.p.m., in 0.10 M HClO_4 saturated with O_2 .

Received 2 February; accepted 29 May 2001.

1. Fan, S. *et al.* Self-oriented regular arrays of carbon nanotubes and their field emission properties. *Science* **283**, 512–514 (1999).
2. Rueckes, T. *et al.* Carbon nanotube-based nonvolatile random access memory for molecular computing. *Science* **289**, 94–97 (2000).
3. Planeix, J. M. *et al.* Applications of carbon nanotubes as supports in heterogeneous catalysis. *J. Am. Chem. Soc.* **116**, 7935–7936 (1994).
4. Rodriguez, N. M., Chambers, A. & Baker, R. T. K. Catalytic engineering of carbon nanostructures. *Langmuir* **11**, 3862–3866 (1995).
5. Dillon, A. C. *et al.* Storage of hydrogen in single-walled carbon nanotubes. *Nature* **386**, 377–379 (1997).
6. Lin, J. Hydrogen storage in nanotubes. *Science* **287**, 1929–1929 (2000).
7. Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego, 1996).
8. Kyotani, T., Tsai, L.-F. & Tomita, A. Formation of ultrafine carbon tubes by using an anodic aluminum oxide film as a template. *Chem. Mater.* **7**, 1427–1428 (2000).
9. Thess, A. *et al.* Crystalline ropes of metallic carbon nanotubes. *Science* **273**, 483–487 (1996).
10. Kruk, M., Jaroniec, M., Ko, C. H. & Ryoo, R. Characterization of the porous structure of SBA-15. *Chem. Mater.* **12**, 1961–1968 (2000).
11. Lettow, J. S. *et al.* Hexagonal to mesocellular foam phase transition in polymer-templated mesoporous silicas. *Langmuir* **16**, 8291–8295 (2000).
12. Ryoo, R., Jun, S., Kim, J. M. & Kim, M. J. Generalised route to the preparation of mesoporous metallosilicates via post-synthetic metal implantation. *Chem. Commun.* 2225–2226 (1997).
13. Jun, S. *et al.* Synthesis of new, nanoporous carbon with hexagonally ordered mesostructure. *J. Am. Chem. Soc.* **122**, 10712–10713 (2000).
14. Ryoo, R., Joo, S. H. & Jun, S. Synthesis of highly ordered carbon molecular sieves via template-mediated structural transformation. *J. Phys. Chem. B* **103**, 7743–7746 (1999).
15. Kruk, M., Jaroniec, M. & Sayari, A. Application of large pore MCM-41 molecular sieves to improve pore size analysis using nitrogen adsorption measurements. *Langmuir* **13**, 6267–6273 (1997).
16. Kageyama, K., Tamazawa, J. & Aida, T. Extrusion polymerization: catalyzed synthesis of crystalline linear polyethylene nanofibers within a mesoporous silica. *Science* **285**, 2113–2115 (1999).
17. Radovic, L. R. & Ridriguez-Reinoso, F. in *Chemistry and Physics of Carbon* Vol. 25 (ed. Thrower, P. A.) 243–358 (Marcel-Dekker, New York, 1997).
18. Kinoshita, K. *Carbon, Electrochemical and Physicochemical Properties* (John Wiley & Sons, New York, 1988).
19. Ryoo, R. *et al.* Application of the xenon-adsorption method for the study of metal cluster formation and growth on Y zeolite. *J. Am. Chem. Soc.* **114**, 76–82 (1992).
20. Kordesch, K. & Simader, G. *Fuel Cells and Their Electrochemistry* (VCH, Weinheim, 1996).
21. Peuckert, M., Yoneda, T., Dalla Betta, R. A. & Boudart, M. Oxygen reduction on small supported platinum particles. *J. Electrochem. Soc.* **133**, 944–947 (1986).
22. Poirier, J. A. & Stoner, G. E. Microstructural effects on electrocatalytic oxygen reduction activity of nano-grained thin-film platinum in acid media. *J. Electrochem. Soc.* **141**, 425–430 (1994).
23. Takasu, Y. *et al.* Size effects of platinum particles on the electroreduction of oxygen. *Electrochim. Acta* **41**, 2595–2600 (1996).
24. Wasmus, S. & Kuver, A. Methanol oxidation and direct methanol fuel cells: a selective review. *J. Electroanal. Chem.* **461**, 14–31 (1999).
25. Yang, P., Zhao, D., Chmelka, B. F. & Stucky, G. D. Triblock-copolymer-directed syntheses of large-pore mesopore silica fibers. *Chem. Mater.* **10**, 2033–2036 (1998).
26. Huang, L. *et al.* Fabrication of ordered porous structures by self-assembly of zeolite nanocrystals. *J. Am. Chem. Soc.* **122**, 3530–3531 (2000).
27. Doshi, D. A. *et al.* Optically defined multifunctional patterning of photosensitive thin-film silica mesophases. *Science* **290**, 107–111 (2000).

28. Fan, H. Y. *et al.* Rapid prototyping of patterned functional nanostructures. *Nature* **405**, 56–60 (2000).
29. Cho, S. J., Ahn, W.-S., Hong, S. B. & Ryoo, R. Investigation of the platinum cluster size and location on zeolite KL with ^{129}Xe NMR, XAFS, and xenon adsorption. *J. Phys. Chem.* **100**, 4996–5003 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

R.R. thanks M. Nomura for helpful discussions on EXAFS measurement. This work was supported in part by the Ministry of Science and Technology through the Creative Research Initiative Program (R.R.), by the School of Molecular Science through the Brain Korea 21 Project (R.R. and J.K.), by the Korea Science and Engineering Foundation through the MICROS Center at KAIST (J.K.), and by CREST, Japan Science and Technology Corporation (O.T.).

Correspondence and requests for materials should be addressed to R.R. (e-mail: rryoo@mail.kaist.ac.kr).

The dating of shallow faults in the Earth's crust

Ben A. van der Pluijm*, Chris M. Hall*, Peter J. Vrolijk†, David R. Pevear† & Michael C. Covey†

* Department of Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA

† ExxonMobil Upstream Research Co., Houston, Texas 77252-2189, USA

Direct dating of ductile shear zones and calculation of uplift/exhumation rates can be done using various radiometric dating techniques. But radiometric dating of shallow crustal faulting, which occurs in the crust's brittle regime, has remained difficult^{1–4} because the low temperatures typical of shallow crusted faults prevent the complete syntectonic mineral recrystallization that occurs in deeper faults. Both old (detrital) and newly grown (authigenic) fine-grained phyllosilicates are thus preserved in shallow fault zones and therefore their radiometric ages reflect a mixture of both mineral populations. Also, the loss of ^{39}Ar during neutron irradiation in dating of clay minerals can produce erroneously old ages. Here we present a method of characterizing the clay populations in fault gouge, using X-ray modelling, combined with sample encapsulation, and show how it can be used to date near-surface fault activity reliably. We examine fault gouge from the Lewis thrust of the southern Canadian Rockies, which we determine to be ~52 Myr old. This result requires the western North America stress regime to have changed from contraction to extension in only a few million years during the Eocene. We also estimate the uplift/exhumation age and sedimentary source of these rocks to be ~172 Myr.

Dating of shallow faults is, among other things, critical for our understanding of crustal evolution, plate interaction and fault reactivation, but there are two obstacles to radiometric dating of clay-rich fault rocks: (1) ^{39}Ar recoil in $^{40}\text{Ar}/^{39}\text{Ar}$ chronology and (2) 'contamination' of samples from old, detrital material. The momentum transfer that occurs during the $^{39}\text{K}(\text{n.p.}) \rightarrow ^{39}\text{Ar}$ reaction is sufficient to move a produced Ar atom about 0.1 μm from the site of the original K atom, which, for clay minerals, can be much greater than the average grain thickness. Thus, one expects massive losses of ^{39}Ar during neutron irradiation, which would lead to erroneously old ages. Vacuum-encapsulated irradiation has been developed as a solution to the recoil problem^{5–8}. The second problem, a mixed age

Table 1 Lewis thrust gouge data

Sample	I in I/S (%)	Detr I (%)	Ar/Ar(total) (Myr)
Bentonitic claystone at fault			
104G-c	70	57	129.6 ± 0.4
104G-m	83	21	81.3 ± 0.4
104G-f	85	12	67.5 ± 0.1
104G-f2	85	12	67.2 ± 0.2
Bentonitic claystone 10 cm from fault			
102E-c	69	73	133.0 ± 0.4
102E-m	80	39	94.6 ± 0.4
102E-f	75	18	72.3 ± 0.1
102E-f2	75	18	72.0 ± 0.3

The table shows percentage of illite in mixed-layer illite/smectite (I/S), percentage of detrital (discrete, 2M₁) illite (Detr I), and total gas Ar ages in Myr (Ar/Ar(total)) for three size fractions of fault gouge samples. Corresponding spectra are shown in Fig. 1.

resulting from the contribution of detrital (old) and newly formed (authigenic) phases, can be resolved through quantitative X-ray analysis of clay grain size populations in low-grade samples. Rather than (erroneously) assuming that little or no detrital material is left in very fine grain size fractions, we quantify the ratio of authigenic and detrital mica in different clay size fractions^{9,10}. This ratio typically decreases with increasing grain size. These grain size fractions are subsequently prepared for Ar dating, which produces a different apparent age for each grain size population. Combined with knowledge of the percentage of detrital illite these apparent ages constrain the age of each end-member phase (that is, of authigenic and detrital clays).

The success of our approach is demonstrated in a suite of gouge samples from the Lewis thrust in the southernmost Canadian Rockies (Gould dome near Crowsnest pass)^{11–13}. This site was selected because faulted mudstone and bentonite units produce excellent outcrops of clay-rich gouge and the geologic age of faulting is reasonably well defined. The oldest age for motion on this fault is defined by the age of the youngest footwall sediments, which are Maastrichtian in age (~65 Myr). The youngest age for thrusting in the area is based on stratigraphic and structural characteristics of early Eocene deposits and is limited by the age of normal faults that

cut the thrust and associated middle Eocene epoch (~48 Myr) deposits^{11–13}. To the south, in the Rocky Mountain foreland of Wyoming, the latest foreland thrusting is also considered to be early Eocene in age^{13,14}.

Three grain size fractions from two sites of the Lewis thrust near Crowsnest pass were prepared¹⁵. The properties of the samples are listed in Table 1 and the corresponding Ar spectra are shown in Fig. 1. Two samples were prepared from the finest grain size fraction and show excellent repeatability. X-ray diffraction analysis shows the Lewis thrust gouge samples to be mixtures of authigenic illite in illite/smectite and discrete detrital illite (mica). Transmission electron microscopy shows that smectite away from the contact is replaced by illite-rich mixed-layer illite/smectite and occasional discrete illite near the contact¹⁶. The Ar data similarly display features that are characteristic of mixed-layer illite/smectite age spectra. Ages start at approximately zero for the room temperature 'recoil' gas fraction, indicating that there was virtually no loss of radiogenic ⁴⁰Ar (ref. 17). Thus, the degree of ³⁹Ar loss due to recoil is based on the sample's structure and is not due to heating from neutron irradiation. After the recoil fraction, ages climb gradually to a level above the total gas age. ³⁹Ar recoil may produce point defects in the clay crystal structure, and is therefore likely to induce enhanced diffusional loss^{18,19}, which accounts for both the rise of ages from zero and an 'overshoot' in apparent ages in what might normally be considered a plateau segment. Plateau ages can therefore only be used with well crystallized (epizonal grade) illite, where the net loss of ³⁹Ar due to recoil is trivial. We also see evidence in the Ar spectra of increasing detrital mica with increasing grain size fraction. Gouge samples show distinctive high age zones at the high-temperature part of the age spectra, which is a feature also noted in synthetic mixtures of clay components⁸ and Gulf coast shale samples¹⁷.

Using modelling of X-ray spectra¹⁸, we determined the percentage of discrete (detrital) illite (of total illite: %detrital + %authigenic = 100%) in each grain size population. Our previous efforts indicate that these estimates have a 1–3% error; in our analysis we have therefore used an average ±2% error. Table 1 lists the data from three size fractions of the two gouge samples. In Fig. 2 we plot percentage detrital illite against the total gas age of the eight analyses

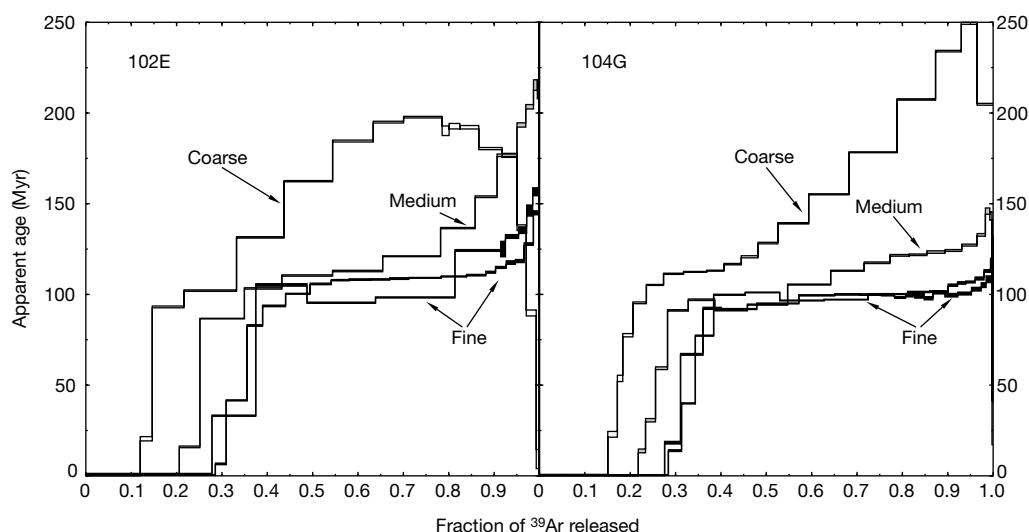


Figure 1 Representative Ar age spectra of clays in samples at the Lewis thrust for three grain size populations. The first fraction in each run is the gas released when the quartz capsule is broken, and represents the gas lost by the sample during neutron irradiation. This 'recoil' gas fraction is always nearly zero in apparent age, meaning that ³⁹Ar is released during irradiation due to recoil, but radiogenic ⁴⁰Ar is not. The amount of recoil ³⁹Ar varies from about 10% to 30% of the total, with the fine-grained samples having a higher percentage loss. This is expected, owing to their higher surface area to volume

ratios, which tends to control the recoil loss mechanism. Apparent ages tend to increase at higher-temperature steps, especially in the coarse-fraction samples. We interpret this as representing the outgassing of relatively well crystallized mica from the host rocks. Fine fractions are <0.02 μm (black boxes), medium are 0.2–0.02 μm (grey boxes) and coarse are 2–0.2 μm (white boxes). Errors are ±1σ. Sample numbers correspond to data in Table 1.

of gouge and the best fit line through these data. The line is an error-weighted least-squares linear regression taking into account measurement errors in both the x and y coordinates. The actual fit was done on the function $\exp(\lambda t) - 1$, which is a linear function of the radiogenic ^{40}Ar to K ratio.

Whereas we observe a large variation in the detrital illite component (ranging from 12–73%) and total gas age (67–133 Myr) in the samples, the results plot along a well defined line with a high degree of precision ($R^2 = 0.96$, mean of squared weighted deviates $\text{MSWD} = 4.8$; Fig. 2). The quoted errors are 1σ and include both *a priori* measurement errors and scatter about the best fit linear regression. Including error estimates for both detrital illite determination and standard Ar analysis error, we derive a lower intercept age at 0% detrital illite of 51.5 ± 3.5 Myr ago (early Eocene), which agrees well with geologic evidence for late movement on the Lewis thrust¹³. The upper intercept of the regression line is calculated as 171.5 ± 6.2 Myr ago, which defines a sample containing 100% detrital material; that is, the 'age' of detrital micas. This middle Jurassic period age represents the mean age of uplift of the source terrain through the $\sim 280^\circ\text{C}$ isotherm, which occurred during exhumation of the internal core of the Cordilleran orogen¹¹. The approximate 52 Myr age of latest contractional faulting in the Canadian Cordillera, combined with geologic evidence for the onset of regional extension soon afterward¹³, requires tectonic processes that allow a dramatic change in stress regime over a period of no more than a few million years. This supports the view that the onset of extension in the Cordillera reflects a change in slab-orogenic lithosphere coupling from delamination or a new subduction geometry²⁰, rather than more gradual deeper mantle or lithospheric weakening processes.

Radiometric dating of near-surface faulting is possible by combined X-ray and Ar analysis of clay separates from fault gouge. X-ray analysis constrains the ratio of authigenic/detrital

material, while modern Ar analysis permits radiometric dating of sub-milligram grain-size fractions. This approach extends reliable dating of crustal deformation to near-surface conditions, which will greatly facilitate the study of crustal evolution and regional tectonics. Because our method also gives the (cooling) age of the source area of the detrital material, it further adds the opportunity to constrain the uplift history and sedimentary source of continental regions. □

Methods

In the vacuum-encapsulation method of Ar dating, the sample is in a fused silica vial that is evacuated to high vacuum and sealed. The capsule is then irradiated in a nuclear reactor and any recoiled ^{39}Ar is trapped within the capsule. In some applications, the whole capsule is fused and the experiment is functionally equivalent to a K–Ar analysis. In others, the capsule is cracked under vacuum so that the recoiled gas could be analysed separately, and the samples are then step-heated. The percentage of recoiled ^{37}Ar (produced by $^{40}\text{Ca}(n, \alpha) \rightarrow ^{37}\text{Ar}$) is equivalent to the percentage of ^{39}Ar released⁶, despite the fact that ^{37}Ar is expected to travel about 2.5 farther than ^{39}Ar recoils, on the basis of conservation of momentum arguments (analogous to illite). We then realized that there is significant redistribution of recoiled Ar atoms from grain to grain, and that nanometre-scale features determine retention of ^{39}Ar . Illite from shales and bentonites of the Welsh Basin and New York State⁷, confirming earlier findings, show an excellent correlation between the illite XRD peak width ($\Delta 2\theta$) and percentage of ^{39}Ar lost due to recoil. The $\Delta 2\theta$ value, called illite crystallinity, is a function of the mean illite diffracting domain thickness (that is, the average number of 1.0-nm illite layers per particle or packet). The advantages of vacuum-encapsulated $^{40}\text{Ar}/^{39}\text{Ar}$ dating over the K–Ar method^{9,10,21,22} are: (1) that it significantly reduces the sample size requirements from ten to hundreds of mg to sample sizes below 1 mg for the $^{40}\text{Ar}/^{39}\text{Ar}$ method; (2) that it avoids possible 'nugget' effects, where the two separate aliquots for K and Ar analysis might not be representative of (sub-)milligram samples, because the $^{40}\text{Ar}/^{39}\text{Ar}$ method measures both radiogenic ^{40}Ar and ^{39}Ar (a proxy for K) on the same sample; and (3) that the precision of analysis for $^{40}\text{Ar}/^{39}\text{Ar}$ is significantly better than for K–Ar methods. Some studies have found that for pure illite or illite/muscovite samples, ages calculated omitting the recoil gas can correct for ^{40}Ar lost owing to structural defects. However, it was demonstrated that for mixed-layer illite/smectite this is an overcorrection¹⁷, and therefore we use the total gas age that includes the recoil gas fraction.

Our Illite Age Analysis (IAA) method capitalizes on the inherently variable ratio of the detrital and authigenic components in different grain size fractions. The detrital mica component is characterized by $2M_1$ polytype, whereas the authigenic form is $1M$ / $1M_d$ polytype (typically mixed-layer illite/smectite) in low-grade shales and mudstones⁸. $2M_1$ mica is considered to be detrital clay as its crystallization temperature exceeds $\sim 280^\circ\text{C}$ (ref. 23). The authigenic/detrital ratio is obtained through iterative modelling of the X-ray diffraction patterns of powdered samples using modified versions of the programs NEWMOD and WILDFIRE^{22,24–26}. Using standard Stoke's Law settling techniques, we separate clay grain size fractions of $2\text{--}0.2\text{ }\mu\text{m}$, $0.2\text{--}0.02\text{ }\mu\text{m}$ and $<0.02\text{ }\mu\text{m}$, from which we determine the authigenic/detrital ratio through X-ray diffraction.

Received 14 June 2000; accepted 19 April 2001.

- Murphy, P. J., Briedis, J. & Peck, J. H. Dating techniques in fault investigations. *Rev. Eng. Geol.* **4**, 153–168 (1979).
- Kralik, M., Klima, K. & Riedmueller, G. Dating fault gouges. *Nature* **327**, 315–317 (1987).
- Gibbons, W. *et al.* Mylonite to megabreccia; tracking fault events within a transcurrent terrane boundary in Nova Scotia, Canada. *Geology* **24**, 411–414 (1996).
- Eide, E. A., Torsvik, T. H. & Andersen, T. B. Absolute dating of brittle fault movements; Late Permian and Late Jurassic extensional fault breccias in western Norway. *Terra Nova* **9**, 135–139 (1997).
- Foland, K. A., Hubacher, F. A. & Arehart, G. B. $^{40}\text{Ar}/^{39}\text{Ar}$ dating of very fine-grained samples: An encapsulated-vial procedure to overcome the problem of ^{39}Ar recoil loss. *Chem. Geol.* **102**, 269–276 (1992).
- Smith, P. E., Evensen, N. M. & York, D. First successful ^{40}Ar – ^{39}Ar dating of glauconites: Argon recoil in single grains of cryptocrystalline material. *Geology* **21**, 41–44 (1993).
- Dong, H., Hall, C. M., Peacor, D. R. & Halliday, A. N. Mechanisms of argon retention in clays revealed by laser ^{40}Ar – ^{39}Ar dating. *Science* **267**, 355–359 (1995).
- Onstott, T. C., Mueller, C., Vrolijk, P. J. & Pevear, D. R. Laser $^{40}\text{Ar}/^{39}\text{Ar}$ microprobe analyses of fine-grained illite. *Geochim. Cosmochim. Acta* **61**, 3851–3861 (1997).
- Peavey, D. R. in *Proc. 7th Int. Symp. on Water-Rock Interactions* (eds Kharaka, Y. K. & Maest, A. S.) 1251–1254 (Balkema, Rotterdam, 1992).
- Pevear, D. R. Illite and hydrocarbon exploration. *Proc. Natl. Acad. Sci.* **96**, 3440–3446 (1999).
- Price, R. A. in *Thrust and Nappe Tectonics* (eds McClay, K. R. & Price, N. J.) 427–448 (Geological Society, London, 1981).
- Fermor, P. Aspect of the three-dimensional structure of the Alberta Foothills and Front Ranges. *Geol. Soc. Am. Bull.* **111**, 317–346 (1999).
- Constenius, K. N. Late Paleogene extensional collapse of the Cordilleran foreland fold and thrust belt. *Geol. Soc. Am. Bull.* **108**, 20–39 (1996).
- Wiltzschko, D. V. & Dorr, J. A. Timing of deformation in the overthrust belt and foreland of Idaho, Wyoming, and Utah. *Am. Ass. Petrol. Geol.* **67**, 1304–1322 (1983).
- Vrolijk, P. & van der Pluijm, B. A. Clay gouge. *J. Struct. Geol.* **21**, 1039–1048 (1999).
- Yan, Y., van der Pluijm, B. A. & Peacor, D. R. Deformation microfabrics of clay gouge, Lewis Thrust, Canada: a case for fault weakening from clay transformation. *Geol. Soc. Spec. Publ.* (in the press).

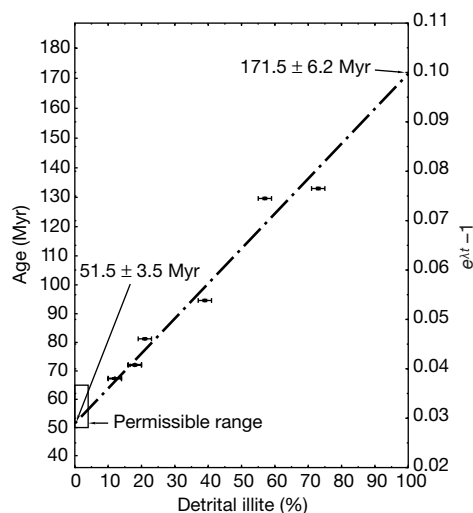


Figure 2 The Illite Age Analysis (IAA) plot correlates the percentage detrital component and age of a sample. The percentage of detrital illite in different grain size fractions is based on X-ray analysis of samples, for which we determine the corresponding Ar total gas ages. The function $e^{\lambda t} - 1$ is linearly proportional to percentage detrital mica and was used to fit the data; λ is decay constant, t is apparent age. The lower intercept of the best-fitting line at 0% detrital illite represents the age of faulting (~ 52 Myr ago), whereas the upper intercept at 100% detrital illite represents the metamorphic/cooling age of micas in the exhumed source region (~ 172 Myr ago). The permissible range of thrust faulting in the sample area based on stratigraphic and cross-cutting relationships is indicated by the box. The age of faulting based on our analysis falls at the young end of this range.

17. Dong, H., Hall, C. M., Peacor, D. R., Halliday, A. N. & Pevear, D. R. Thermal $^{40}\text{Ar}/^{39}\text{Ar}$ separation of diagenetic from detrital illitic clays in Gulf Coast shales. *Earth Planet. Sci. Lett.* **175**, 309–325 (2000).
18. Hall, C. M. *et al.* Dating of alteration episodes related to mercury mineralization in the Almadén district, Spain. *Earth Planet. Sci. Lett.* **148**, 287–298 (1997).
19. Jaboyedoff, M. & Cosca, M. A. Dating incipient metamorphism using $^{40}\text{Ar}/^{39}\text{Ar}$ geochronology and XRD modeling: a case study from the Swiss Alps. *Contrib. Mineral. Petrol.* **135**, 93–113 (1999).
20. Bird, P. Formation of the Rocky Mountains, western United States: a continuum computer model. *Science* **239**, 1501–1507 (1988).
21. Vrolijk, P., Covey, M. C., Pevear, D. R. & Longstaffe, F. Dating clay-rich thrust faults. *Geol. Soc. Am. (Abstr. Progr.)* **26**, 466 (1994).
22. Pevear, D. R. & Schuette, J. F. in *Computer Applications to X-ray Diffraction Analysis of Clay Minerals* (eds Reynolds, R. C. & Walker, J. R.) 19–42 (Clay Minerals Society, Boulder, CO, 1993).
23. Srodon, J. & Eberl, D. D. *Review in Mineralogy* (ed. Bailey, S. W.) 495–544 (Mineralogical Society of America, Washington, DC, 1984).
24. Grathoff, G. H. & Moore, D. M. Illite polytype quantification using Wildfire calculated X-ray diffraction patterns. *Clay, Clay Mineral.* **44**, 835–842 (1996).
25. Reynolds, R. C. WILDFIRE: A computer program for the calculation of three-dimensional X-ray diffraction patterns for mica polytypes and their disordered variations (Hanover, New Hampshire, 1994).
26. Reynolds, R. C. & Reynolds, R. C. NEWMOD: A computer program for the calculation of one-dimensional diffraction patterns of mixed-layered clays. (Hanover, New Hampshire, 1996).

Acknowledgements

D. R. Pevear has retired from ExxonMobil Upstream Research Company. We thank D. R. Peacor for assistance and several Cordilleran geologists for discussion, and the National Science Foundation and ExxonMobil Upstream Research Company for support of our fault gouge research.

Correspondence and requests for materials should be addressed to B.v.d.P. (e-mail: vdpluijm@umich.edu).

Geology and palaeontology of the Late Miocene Middle Awash valley, Afar rift, Ethiopia

Giday WoldeGabriel*, Yohannes Haile-Selassie†, Paul R. Renne‡, William K. Hart§, Stanley H. Ambrose||, Berhane Asfaw‡, Grant Heiken# & Tim White†

* EES-6/MS D462; and # Institute of Geophysics and Planetary Physics, MS C303, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

† Department of Integrative Biology and Laboratory for Human Evolutionary Studies, Museum of Vertebrate Zoology, 3060 VLSB, University of California, Berkeley, California 94720, USA

‡ Berkeley Geochronology Center, 2455 Ridge Road, and Department of Earth and Planetary Science, University of California, Berkeley, California 94709, USA

§ Department of Geology, Miami University, Oxford, Ohio 45056, USA

|| Department of Anthropology, University of Illinois, Urbana, Illinois 61801, USA

¶ Rift Valley Research Service, P. O. Box 5717, Addis Ababa, Ethiopia

The Middle Awash study area of Ethiopia's Afar rift has yielded abundant vertebrate fossils ($\approx 10,000$), including several hominid taxa^{1–4}. The study area contains a long sedimentary record spanning Late Miocene (5.3–11.2 Myr ago) to Holocene times. Exposed in a unique tectonic and volcanic transition zone between the main Ethiopian rift (MER) and the Afar rift, sediments along the western Afar rift margin in the Middle Awash provide a unique window on the Late Miocene of Ethiopia. These deposits have now yielded the earliest hominids, described in an accompanying paper⁵ and dated here to between 5.54 and 5.77 Myr. These geological and palaeobiological data from the Middle Awash provide fresh perspectives on hominid origins and early evolution. Here we show that these earliest hominids derive from relatively wet and wooded environments that were modulated

by tectonic, volcanic, climatic and geomorphic processes. A similar wooded habitat also has been suggested for the 6.0 Myr hominoid fossils recently recovered from Lukeino, Kenya⁶. These findings require fundamental reassessment of models that invoke a significant role for global climatic change and/or savannah habitat in the origin of hominids.

The western rift margin is more than 30-km wide, and drops in elevation from greater than 2,500 m on the plateau to about 600 m at the rift floor. It is attenuated, with east-dipping, distinct arcuate antithetic morphology from fault displacement in a tectonic transfer zone between the NNW- and NNE-trending Red Sea and MER tectonic domains, respectively^{7,8} (Fig. 1, inset). Zones of broad warping along rift margins are typical of transfer zones in extensional regions such as the east African rift system⁷. The transfer zone is permeated by dike swarms⁹, and such magma flux and dike injection along steep boundary faults during rifting probably increased geothermal gradient, ductile deformation and crustal separation in the southern Afar rift margin. The close association between rifting and development of transfer zones exerts significant influence on structural patterns and synrift sedimentation⁷. The

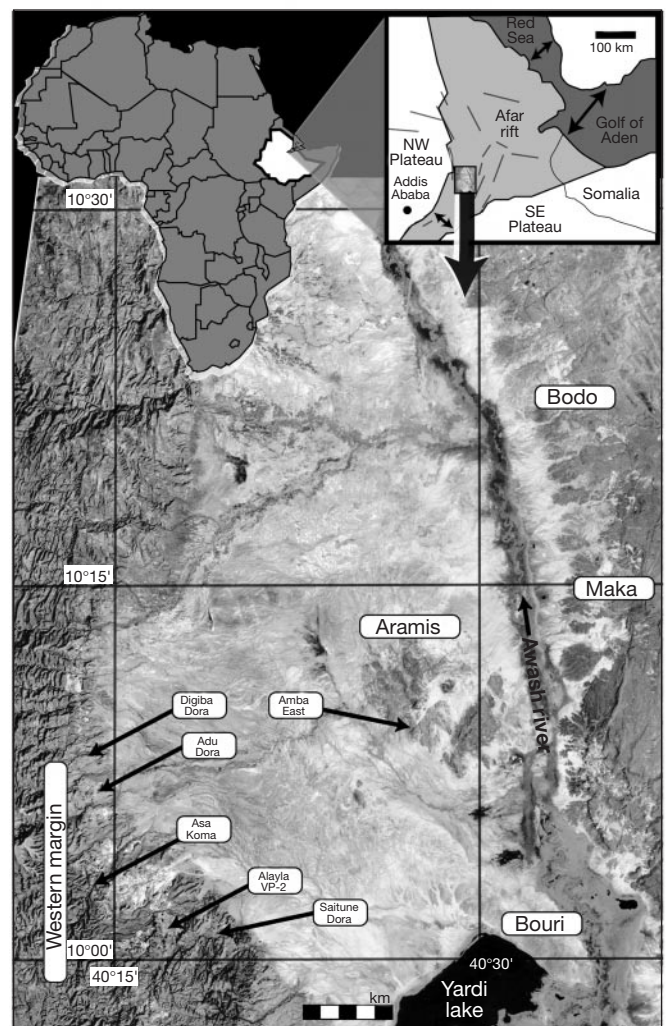
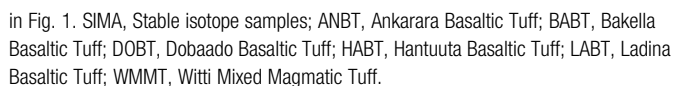


Figure 1 Location map showing measured sections along the western rift margin of the Middle Awash region of the southern Afar rift. Map based on Landsat Thematic Mapper imagery. Complex linear and arcuate NE-trending and transverse faulting is apparent along the rift margin. The broad rift margin and rift floor are shown by darker and lighter shades, respectively. Other hominid sites within the Middle Awash study area are located at Aramis (4.4 Myr; *Ardipithecus ramidus*), Maka (3.4 Myr; *Australopithecus afarensis*), Bouri (2.5 Myr; *Australopithecus garhi*) and Bodo (0.64 Myr; *Homo*).

The older Adu-Asa Formation was originally divided into the Adu and Asa Members¹². Subsequent work required that each of these former members be divided and the resulting four members be defined on the basis of associated tephra markers and distinctive lithofacies¹⁵. We divide the former Adu Member into the fluvial Saraitu and overlying lacustrine Adu Dora Members. The type section for the Saraitu Member is in the Saraitu drainage 3 km north of Alayla (Fig. 1). Here, and at Gaisale, Asa Koma and Adu Dora North, ≤ 50 m of variegated fluvial deposits and thin (3–5 cm) basaltic and silicic tephra are exposed (Fig. 2). Basaltic lavas yielding incremental heating plateau ages of 6.33 ± 0.07 Myr (MA95-1) and 6.16 ± 0.06 Myr (MA95-22) underlie the Saraitu Member. These dates provide a maximum age for the Adu-Asa

Geochronology of the hominid-bearing Asa Koma Member is firmly grounded by the 5.77 ± 0.08 Myr old Ladina Basaltic Tuff (LABT, MA97-15) near the base of the member. Higher in the member is the Wittl Mixed Magmatic Tuff (WMMT). Splits of basaltic glass in this tuff from Asa Koma (MA96-30) yielded indistinguishable plateau ages of 5.63 ± 0.12 and 5.57 ± 0.08 Myr,



whereas plagioclase phenocrysts from the correlative MA95-4 at Digiba Dora yielded a weighted (by inverse variance) mean age of 5.68 ± 0.07 Myr for 45 individually analysed crystals. At Saitune Dora, the VP-1 and VP-2 fossil localities are capped by basalt (MA95-7), yielding a plateau age of 5.54 ± 0.17 Myr. Therefore, all of the vertebrate fossils discussed here are firmly bracketed between 5.54 and 5.77 Myr. Palaeomagnetic samples collected in this interval (Fig. 2) yield uniformly reverse polarity, as expected from the geomagnetic polarity timescale¹⁶.

Waning phreatomagmatic eruptions of the Asa Koma Member marked the local transition from the Miocene to the Pliocene¹⁷. The post-phreatomagmatic sequence of the Rawa Member at the Asa Koma type section is ≈ 75 -m thick and consists of conglomerate, reddish brown silty clays and palaeosols. The Hantuuta Basaltic Tuff (HABT, MA98-45) marks the base of an overlying, distinct lithologic assemblage that defines the top of the Rawa Member. The sequence above the Rawa Member at the Asa Koma section contains silicic ashes and welded ignimbrite, which are widely distributed along strike of the frontal fault blocks. They are interbedded with widespread pedogenic carbonates (≥ 5 -m thick), palaeosols and coarse clastic deposits, reflecting changes in sedimentation processes and environmental conditions. A fifth member may be designated for the Adu-Asa Formation when the detailed investigation of the post-Rawa Member sequence is completed.

The succession described here records tectonic, volcanic and sedimentary processes along the western margin and adjacent basin floor during the Late Miocene. Local and regional tephra correlations of isolated sections measured in the densely step-faulted antithetic blocks of the escarpment allowed characterization of the local stratigraphy (Fig. 2; see also Supplementary Information for microprobe glass chemistry data from marker tephra).

The earliest hominid remains in Ethiopia have been discovered among more than 60 identified vertebrate species on the basis of over 1,900 fossils from the Middle Awash. Approximately 15% of this collection comes from the Kuseralee Member of the Sagantole Formation of the CAC (> 5.2 Myr¹⁴). The older fossils (5.54–5.77 Myr) from the Asa Koma Member are primarily derived from fluviatile deposits. Biochronological assessment is consistent with the isotopic dating. Biochronologically important mammals include the suid *Nyanzachoerus syrticus* and the very primitive *Primelephas*.

The thick fluvial lacustrine and phreatomagmatic units of the lower Adu-Asa Formation record environmental conditions much wetter than today. This is consistent with post-rifting depositional profiles from the Gulf of Suez and seven locations along the western coastal areas of the Red Sea¹⁸, Late Miocene to Middle Pliocene (3.4 Myr) fossil floral and isotopic evidence^{19,20}, and low-resolution pollen data from Ocean Drilling Program (ODP) site 721 (refs 21, 22). There is strong potential for linking the Middle Awash record with marine homologues using silicic tephra recovered from Red Sea, Gulf of Aden, and the western Indian Ocean cores to understand the Neogene geological and palaeoenvironmental processes of the adjacent, evolving continental and oceanic rifts.

Low carbonate carbon isotope ratios indicate woodland to grassy woodland²³ (20–45% grass biomass) habitats at the Asa Koma and Digiba Dora hominid sites. Low oxygen isotope ratios indicate cool, high altitude and/or humid habitats²⁴ (see Supplementary Information for stable isotope information).

Additional palaeoenvironmental indicators are vertebrate fossils from the four western margin localities that have yielded ten hominid specimens. These fossil assemblages indicate predominantly wet and closed woodland/forest habitats. The abundance of reduncine bovids also indicates the presence of open woodland or wooded grassland around lake margins. Among the micromammals, *Tachyoryctes* (root rats) and *Thryonomys* (cane rat) are abundant. Extant *Thryonomys* species are known to live along margins of rivers and lakes²⁵. Extant *Tachyoryctes* species live in

highland settings, consistent with the Adu-Asa Formation being deposited at a higher elevation during the end of the Miocene. Alcelaphines are absent in the formation, in contrast with their abundance in the Nawata Formation of Lothagam, Kenya. The rarity of lagomorphs (hares; only one species present), shows that open grassland habitats are not well sampled in the fauna of the Adu-Asa Formation.

The dominant fauna from the younger Kuseralee Member of the CAC, 20 km to the east, indicates more open woodland and/or lake marginal habitat. Rich vertebrate assemblages are known from these more easterly, axial basin depositories. These sediments have yielded abundant fossils of *Anancus* and *Nyanzachoerus*. However, only one hominid specimen has been recovered from Kuseralee Member deposits despite intensive searching. This specimen from Amba East was found associated within a stratigraphically restricted, more subaerial deposit containing carnivores, bovids and cercopithecoid monkeys. Isotopic results suggest a warmer, lower altitude and/or drier grassy woodland to woodland floral habitat for the hominid-associated, stratigraphically restricted CAC assemblage.

The end Miocene fauna from the Asa Koma Member is roughly contemporary with the Lothagam fauna from the upper Nawata Formation of Kenya, where a mosaic of habitats was present near the Miocene/Pliocene boundary²⁶. However, hominids are nearly absent from the Nawata Formation. The demonstration that the earliest hominids consistently derive from strata bearing indicators of wooded environments may explain their rarity at some sites. It therefore seems increasingly likely that early hominids did not frequent open habitats until after 4.4 Myr¹³. Before that, they may have been confined to woodland and forest habitats. □

Received 19 February; accepted 15 May 2001.

- Conroy, G. C., Jolly, C. J., Cramer, D. & Kalb, J. E. Newly discovered fossil hominid skull from the Afar Depression, Ethiopia. *Nature* **275**, 67–70 (1978).
- White, T. D. et al. New discoveries of *Australopithecus* at Maka in Ethiopia. *Nature* **366**, 261–265 (1993).
- White, T. D., Suwa, G. & Asfaw, B. *Australopithecus ramidus*, a new species of early hominid from Aramis, Ethiopia. *Nature* **371**, 306–312 (1994).
- Asfaw, B. et al. *Australopithecus garhi*: a new species of early hominid from Ethiopia. *Science* **284**, 625–629 (1999).
- Haile-Selassie, Y. Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412**, 178–181 (2001).
- Pickford, M. & Senut, B. The geological and faunal context of Late Miocene hominid remains from Lukeino, Kenya. *C.R. Acad. Sci. Ser. IIa* **332**, 145–152 (2001).
- Morley, C. K., Nelson, R. A., Patton, T. L. & Munn, S. G. Transfer zones in the East African Rift System and their relevance to hydrocarbon exploration in rifts. *Am. Assoc. Petrol. Geol.* **74**, 1234–1253 (1990).
- Morton, W. H. & Black, R. in *Afar Depression of Ethiopia* (eds Pilger, A. & Rosler, A.) 55–61 (Schweizerbart, Stuttgart, 1975).
- Mohr, P. The Morton-Black hypothesis for the thinning of continental crust-revisited in western Afar. *Tectonophysics* **94**, 509–528 (1983).
- Tiercelin, J. J., Taieb, M. & Faure, H. Continental sedimentary basins and volcano-tectonic evolution of the Afar Rift. *Atti Convegni Lincei* **47**, 491–504 (1980).
- Kalb, J. E., Oswald, E. B., Mebrate, A., Tebedge, S. & Jolly, C. J. Stratigraphy of the Awash Group, Middle Awash Valley, Afar, Ethiopia. *Newsl. Stratigr.* **11**, 95–127 (1982).
- Kalb, J. E. Refined stratigraphy of the hominid-bearing Awash Group, Middle Awash Valley, Afar depression, Ethiopia. *Newsl. Stratigr.* **29**, 21–62 (1993).
- WoldeGabriel, G. et al. Ecological and temporal placement of early Pliocene hominids at Aramis, Ethiopia. *Nature* **371**, 330–333 (1994).
- Renne, P. R., WoldeGabriel, G., Hart, W. K., Heiken, G. & White, T. D. Chronostratigraphy of the Miocene-Pliocene Sagantole Formation, Middle Awash Valley, Afar Rift, Ethiopia. *Geol. Soc. Am. Bull.* **111**, 869–885 (1999).
- Hedberg, H. D. *International Stratigraphic Guide* (John Wiley and Sons, New York, 1976).
- Cande, S. C. & Kent, D. V. Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **B 100**, 6093–6095 (1995).
- Harland, W. B. et al. *A Geologic Time Scale* (Cambridge Univ. Press, Oxford, 1990).
- Griffin, D. L. The late Miocene climate of northeastern Africa: unraveling the signals in the sedimentary succession. *J. Geol. Soc. Lond.* **156**, 817–826 (1999).
- deMenocal, P. B. Plio-Pleistocene African climate. *Science* **270**, 53–59 (1995).
- Yemane, K., Bonnefille, R. & Faure, H. Palaeoclimatic and tectonic implications of Neogene microflora from the Northwestern Ethiopian highlands. *Nature* **318**, 653–656 (1985).
- Van Campo, E. Pollen transport into Arabian Sea sediments. *Proc. ODP* **117**, 277–280 (1991).
- Ruddiman, W. F. et al. Late Miocene to Pleistocene evolution of climate in Africa and the low-latitude Atlantic: overview of Leg 108 results. *Proc. ODP* **108**, 463–484 (1989).
- Cerling, T. E., Quade, J., Wang, Y. & Bowman, J. R. Carbon isotopes in soils and paleosols as ecology and paleoecology indicators. *Nature* **341**, 138–139 (1989).
- Quade, J., Cerling, T. E. & Bowman, J. R. Systematic variations in the carbon and oxygen isotopic

composition of pedogenic carbonate along elevation transects in the southern Great Basin, United States. *Geol. Soc. Am. Bull.* **101**, 464–475 (1989).

25. Kingdon, J. *East African Mammals* Vol. IIB (Academic, New York, 1974).

26. Leakey, M. G. *et al.* Lothagam: a record of faunal change in the Late Miocene of East Africa. *J. Vert. Paleontol.* **16**, 556–570 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

The Middle Awash Project is multinational, interdisciplinary research co-directed by B.A., Y. Beyene, J. D. Clark, T. D.W. and G.W.G. The research was supported by the National Science Foundation and the Institute of Geophysics and Planetary Physics of the University of California at Los Alamos National Laboratory. Additional contributions were made by the Graduate School, the Office for Advancement of Scholarship and Teaching and the Geology Department at Miami University, and the Research Board of the University of Illinois. We thank H. Gilbert for field and illustrations work; H. Saegusa for proboscidean identifications; D. DeGusta for primate identifications; F. C. Howell for carnivore identifications; H. Wesselman and M. Asnake for micromammal analysis and identifications; E. Vrba for bovid identifications; and L. Smeenk for palaeomagnetic analyses. We thank the Ministry of Information and Culture, the Authority for Research and Conservation of the Cultural Heritage, and the National Museum of Ethiopia for permission to conduct the research. We appreciate the support of the Afar regional government and the Afar people of the Middle Awash. Access to the Electron Microprobe Laboratory and additional support from the Earth Environmental Sciences Division, Los Alamos National Laboratory, and help from P. Snow is greatly appreciated. S. Baldrige internally reviewed the manuscript at Los Alamos.

Correspondence and requests for materials should be addressed to G.W.G. (e-mail: wgday@lanl.gov).

Late Miocene hominids from the Middle Awash, Ethiopia

Yohannes Haile-Selassie

Department of Integrative Biology and Laboratory for Human Evolutionary Studies, Museum of Vertebrate Zoology, 3060 VLSB, University of California, Berkeley, California 94720, USA

Molecular studies suggest that the lineages leading to humans and chimpanzees diverged approximately 6.5–5.5 million years (Myr) ago, in the Late Miocene^{1–3}. Hominid fossils from this interval, however, are fragmentary and of uncertain phylogenetic status, age, or both^{4–6}. Here I report new hominid specimens from the Middle Awash area of Ethiopia that date to 5.2–5.8 Myr and are associated with a wooded palaeoenvironment⁷. These Late Miocene fossils are assigned to the hominid genus *Ardipithecus* and represent the earliest definitive evidence of the hominid clade. Derived dental characters are shared exclusively with all younger hominids. This indicates that the fossils probably represent a hominid taxon that postdated the divergence of lineages leading to modern chimpanzees and humans. However, the persistence of primitive dental and postcranial characters in these new fossils indicates that *Ardipithecus* was phylogenetically close to the common ancestor of chimpanzees and humans. These new findings raise additional questions about the claimed hominid status of *Orrorin tugenensis*⁸, recently described from Kenya and dated to ~6 Myr⁹.

The western margin of the Middle Awash contains predominantly Late Miocene sediments mostly pre-dating the Kuseralee Member at the base of the Sagantole Formation of the Central Awash Complex (CAC)¹⁰. Palaeontological work since 1992 has yielded abundant vertebrate fossils, including hominids that date to 5.2–5.8 Myr (Table 1). Environmental indicators suggest a wooded habitat⁷. To date, 11 hominid specimens (Fig. 1) have been recovered at five localities since the first (a partial mandible) was recovered from Alayla in 1997. They represent at least five individuals and

seem to represent a single taxon, a new subspecies of *Ardipithecus* (see Methods).

The first specimen recovered was the subspecific holotype, ALA-VP-2/10, a right mandible with M₃. (Note that subscripts indicate lower teeth, superscripts upper teeth.) Four isolated left lower teeth (I₂, L_C, P₄ and M₂) are associated by spatial proximity, colour, perimortem root fracture and wear. The left I₂ is metrically and morphologically comparable to known later hominid incisors and distinctively narrower than the lateral incisors of chimpanzees (*Pan troglodytes*). The P₄ has a well developed talonid and a Tome's root rather than the single roots reported for Aramis A. *ramidus*¹¹.

The associated lower canine is worn apically and distally. Its mesial crown shoulder is elevated relative to the condition usually seen in modern female apes. A distinct marginal ridge is formed on the mesial lingual face. Its distal face has an exposed dentine strip from apex to distal tubercle. The large distal tubercle is shared with Aramis homologues, but the posterior orientation of the wear facet is also shared by apes with a honing canine–premolar complex. However, the distal tubercle in apes is usually worn diagonally as the upper canine extends in full occlusion below the cervico-enamel junction of the lower canine. The distal tubercle in *Ardipithecus* is worn horizontally. The functional implication of this distinction is a possible absence of a fully functional honing canine–premolar complex in *Ardipithecus*.

The M₃ shows small occlusal wear facets on the buccal slopes of the spiky metaconid and entoconid. The buccal cusps are highly worn, with a deep, cupped, coalesced dentine exposure centred at the protoconid. The M₃ of ARA-VP-1/128 (*A. ramidus*) shows a different wear pattern in which both protoconid and metaconid exhibit small apical perforations in the enamel. All later hominids have cusps that are more rounded before wear. The ALA-VP-2/10 and ARA-VP-1/128 lower third molars are similar in mesiodistal dimension. However, ALA-VP-2/10 is absolutely smaller than the known ranges of *A. anamensis* (*n* = 5) and *A. afarensis* (*n* = 14), and absolutely larger than homologues in a sample of 20 common chimpanzees. The M₂ displays a buccal occlusal half deeply excavated by wear, with a large, oval, cupped dentine exposure spanning the protoconid and hypoconid and a separate deep, round exposure at the hypoconulid position. As with the M₃, this wear pattern is different from that of later hominids owing to the extreme wear differential between the lingual and buccal cusps.

A periodontal abscess affects the P₄/M₁ area, and consequent lateral corpus swelling resulted in only slight hollowing from P₄ to posterior M₁. The submandibular fossa is shallow anteriorly. The circular, anterosuperiorly opening mental foramen is positioned at or mesial to P₄ at approximately midcorpus. The preserved corpus is comparable in absolute size to AL 288-1 (*Australopithecus afarensis*) but is less robust at the M₂ and M₃ levels than AL 288-1 or KNM-LT 329 (the Lothagam mandible).

ASK-VP-3/160 is a left P³ crown at an early wear stage. The root is entirely missing. The occlusal crown morphology is similar to Aramis homologues, but the mesial fovea is shallower. In mesial aspect, the mesial marginal ridge of ASK-VP-3/160 is below mid-crown level. Its lingual extension bears an occlusal facet suggesting a prominent P₃ protoconid. It lacks the strong mesiobuccal crown extension commonly seen in *Pan* P³ teeth.

STD-VP-2/61 is a narrow, pointed, unworn lower right canine with three strong horizontal buccal hypoplastic lines. The distal tubercle is less prominent than on ALA-VP-2/10. The mesial crown shoulder is lower (at midcrown) than the contemporary Alayla lower canine. One morphological feature that this canine shares with chimpanzees rather than later hominids is the flattening of the mesiolingual face with an absence of a distinct marginal ridge defined by a vertical mesiolingual groove. The weak development of later hominid lower canine traits on STD-VP-2/61, as well as the tall, narrow apex, makes this the most primitive hominid canine yet found.

STD-VP-2/62 is a fully erupted and minimally worn LM³. The protocone apex bears a small wear facet. Occlusal outline is a buccolingually elongated rectangle with the distal half slightly buccolingually shorter than the mesial. With all four cusps well defined, this molar does not show the noticeable distal tapering usually seen in later hominids, and the occlusal surface is less crenulated than in chimpanzees. The mesial fovea is shallow and not as broad as in chimpanzees. The specimen is similar in size to the reported Aramis M³ (ref. 11).

STD-VP-2/63 is LM¹ with both protocone and paracone exhibiting deeply pitted dentine exposures. It is absolutely smaller than known *A. afarensis* M¹ teeth. It is differentiated from chimpanzee M¹ teeth by the absence of strong occlusal crenulation.

The teeth of these Late Miocene *Ardipithecus* specimens show a mosaic of primitive and derived morphological features. Studies of enamel thickness are underway, but the available broken and little-worn teeth suggest that molar enamel thicknesses in the STD and ALA hominids were comparable to, or slightly greater than, those of the younger Aramis samples of *A. ramidus*. The presence of four distinct cusps and the absence of the distal tapering of the M³ are primitive features shared with most Miocene hominoids. The lower canines are of particular interest. The development of the distal tubercle on these new *Ardipithecus* lower canines and the observed variation in the position of the mesial crown shoulder and expression of the mesial marginal ridge are best interpreted as representing early manifestations of the evolution of an incisiform canine, a

definitive feature of later hominids.

ALA-VP-2/11 is the distal half of an intermediate hand phalanx. Dorsal shaft curvature is minimal. The concave palmar surface is marked by deep, bilateral fossae for the *m. flexor digitorum superficialis*. It is larger than but morphologically similar to most *A. afarensis* intermediate hand phalanges. Head diameter is larger than the largest *A. afarensis* intermediate hand phalanx (AL 333x-46) and comparison indicates that it was very probably longer than the longest *A. afarensis* homologue. DID-VP-1/80 is the distal half of a proximal hand phalanx. Ridges for the *m. flexor retinaculum* are not as developed as in most *A. afarensis* specimens. The overall degree of curvature of DID-VP-1/80 is similar to that of *A. afarensis*.

ASK-VP-3/78 is a left distal humerus fragment preserving some of the trochlea, the base of the medial epicondyle, the olecranon fossa and part of the distal shaft. The medial aspect of the proximal edge of the trochlear joint surface shows post-mortem subchondral erosion and minor arthritic lipping. The specimen is slightly smaller than ARA-VP-7/2 (*A. ramidus ramidus*)¹¹ and absolutely larger than small *A. afarensis* specimens such as AL 288-1m and AL 322-1. Radial and coronoid fossae are separated by a prominent ridge (but not by a 'Herskovitz' tubercle). ASK-VP-3/78 is similar to ARA-VP-7/2 in having a relatively sharp lateral trochlear crest (as in most modern apes and some *A. afarensis*). The olecranon fossa of ASK-VP-3/78 differs from later hominids, which have more elliptical and shallower fossae.

ALA-VP-2/101 is an associated humeral mid-shaft and proximal

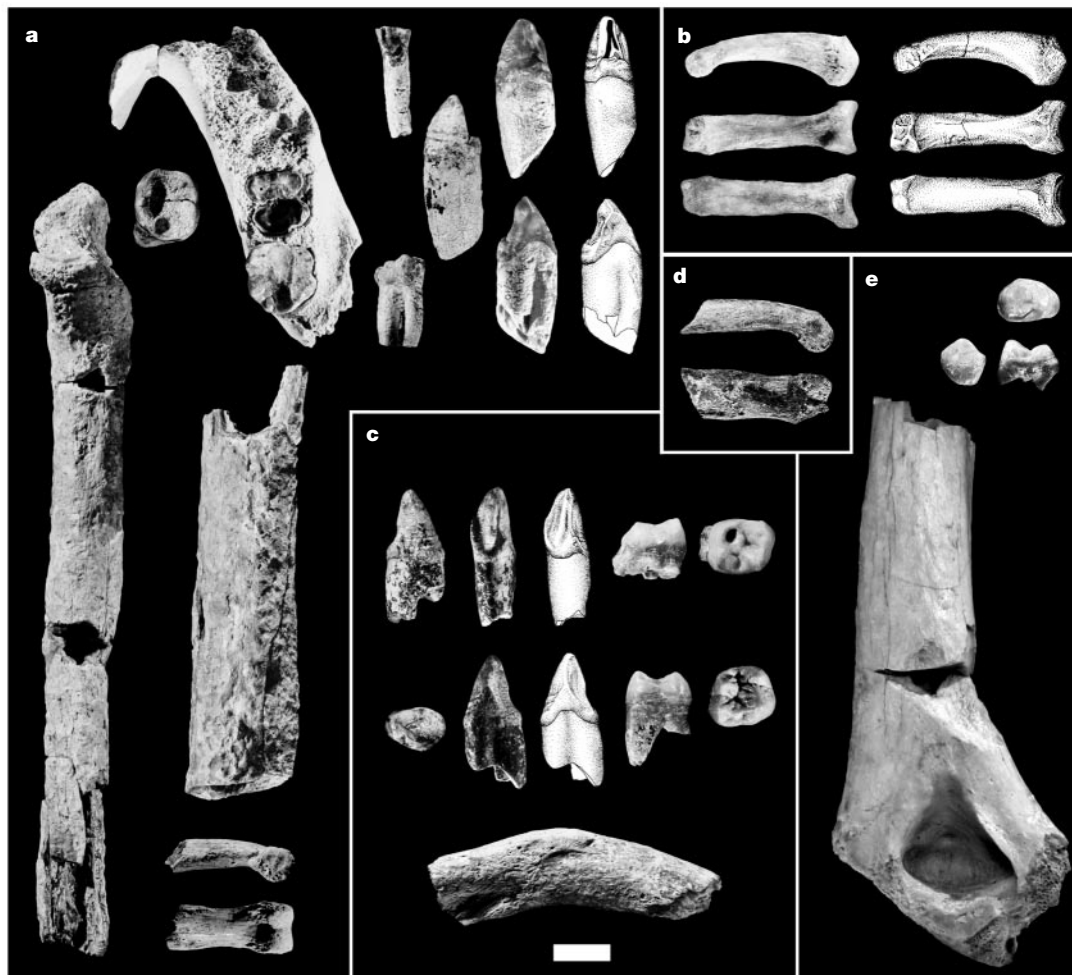


Figure 1 Fossil hominid remains from the Late Miocene Middle Awash deposits. **a**, ALA-VP-2/10, mandible and all associated teeth; ALA-VP-2/120, ulna and humerus shaft; ALA-VP-2/11, hand phalanx. **b**, AME-VP-1/71, lateral, plantar and dorsal views of foot

phalanx. **c**, STD-VP-2, teeth and partial clavicle. **d**, DID-VP-1/80, hand phalanx. **e**, ASK-VP-3/160, occlusal, mesial and buccal views; ASK-VP-3/78, posterior view. All images are at the same scale. Scale bar, 1 cm. Line drawings by L. Gudz.

Table 1 Fossil specimens of *Ardipithecus ramidus kadabba*

Specimen no.	Year collected	Element	Discoverer	Dental dimensions (mm)
AME-VP-1/71	1999	Proximal foot phalanx	L. Hlusko	
ALA-VP-2/10	1997–99	Right mandible with M ₃ and associated teeth	Y. Haile-Selassie	RM ₃ (13.3)MD; LI ₂ 6.3MD, 8.3LL; L _C 11.2MD, 7.8LL, (13.1+)CH; LP ₄ (8.1)MD, 10.0BL; LM ₂ (12.7)MD, 11.8BL
ALA-VP-2/11	1997	Intermediate hand phalanx fragment	S. Eshete	
ALA-VP-2/101	1999	Left humerus and ulna	T. White	
ASK-VP-3/78	1998	Left distal humerus	Y. Haile-Selassie	
ASK-VP-3/160	2001	LP ³	Group	7.6MD, 11.3BL
DID-VP-1/80	1998	Proximal hand phalanx fragment	Y. Haile-Selassie	
STD-VP-2/61	1998	R _C	M. Humed	10.8MD, 7.8LL, 14.3CH
STD-VP-2/62	1998	LM ³	Y. Haile-Selassie	10.9MD, 12.2BL
STD-VP-2/63	1999	LM ¹	Group	(10.6)MD, 12.1BL
STD-VP-2/893	1998	Left clavicle fragment	Y. Haile-Selassie	

Numbers in parentheses are estimates. BL, buccolingual; LL, labiolingual; MD, mesiodistal; CH, crown height.

ulna. The humeral mid-shaft is smaller than that of ARA-VP-7/2 and matches the smallest *A. afarensis*. The ulna is more complete. Most of its shaft is preserved but abraded. The coronoid and olecranon processes and the radial facet are damaged. The insertion area for the *brachialis* muscle is neither excavated nor medially or laterally well marked. Despite being incomplete, the ulnar shaft appears more curved than is typical of most later hominids.

The chronologically younger (5.2 Myr) AME-VP-1/71 is a complete left fourth proximal foot phalanx with a maximum length of 31.9 mm. It is close in maximum length to AL 333-71 (32.5 mm) (ref. 12). In lateral view, the shaft shows strong plantar curvature also comparable to AL 333-71. The distal half of the shaft is dorsoventrally compressed, whereas the proximal half is mediolaterally compressed with a prominent constriction above the base. AME-VP-1/71 shows a mosaic of features shared with both apes and *A. afarensis*. The proximal pedal phalanges of *A. afarensis* are unique in combining both strong phalangeal curvature (similar to apes) with a dorsally canted proximal joint surface (similar to later hominids)¹³. The dorsal orientation of this surface in AME-VP-1/71 may therefore constitute important evidence of a unique pedal morphology in this specimen similar to that in Hadar.

STD-VP-2/893 is the lateral half of a left clavicle lacking the acromial extremity. The deltoid muscle attachment is well marked on the superior surface. The shaft cortex is thick, with an oval cross-section immediately medial to the deltoid attachment. The conoid tubercle is a mediolaterally elongate, roughened surface comparable in overall robustness to AL 333x-9 and absolutely more robust than in chimpanzees.

The Middle Awash fossils described above share some dental characters exclusively with later hominids, and do so to the exclusion of all fossil and extant apes. These characters include lower canines with developed distal tubercles and expressed mesial marginal ridges. In addition, the proximal foot phalanx from Amba, dated at 5.2 Myr, is derived relative to all known apes and is consistent with an early form of terrestrial bipedality. Because of this combination of characters, the Middle Awash fossils described here are classified as cladistically hominid. They are currently distinguishable from the later *A. ramidus* at the subspecies level by more primitive dental characters consistent with their antiquity (see Methods). However, larger samples may reveal additional evidence that will require elevation of this subspecies to species rank.

Another candidate for hominid ancestry is the recently described *Orrorin tugenensis*⁸. The authors report thick molar enamel and suggest that *Ardipithecus* and African apes are commonly derived in having 'thin' enamel. However, enamel thickness is a complex character and intraspecifically variable, and its within-tooth three-dimensional patterning is characteristically expressed both serially and taxonomically. Therefore, the simplistic dichotomous characterization of enamel as either 'thick' or 'thin' on the basis of

unspecified measurements of naturally broken sections (as was done in the *Orrorin* report⁸) is problematic.

The upper canine morphology of *O. tugenensis* is quite primitive, as it lacks the derived, elevated crown shoulders shared by *Ardipithecus* and all other hominids. Furthermore, the locomotor anatomy of *Orrorin* remains uncertain at this time because its description lacked comment on characters directly diagnostic of bipedality, such as the presence of an obturator externus groove¹⁴ or an asymmetrical distribution of cortex in the femoral neck¹⁵. Given its antiquity and characters, as currently described, there is nothing to preclude *Orrorin* from representing the last common ancestor, and thereby antedating the cladogenesis of hominids. It is equally plausible that it represents a previously unknown African hominoid with no living descendants, or an exclusive precursor of chimpanzees, gorillas or humans.

The phylogeny proffered in the description of the *Orrorin* fossils interprets *Ardipithecus* as a chimpanzee ancestor⁸. The authors state that this view is consistent with early hominids evolving east of the African rift system and chimpanzees and gorillas evolving to the west. But how could a putative chimpanzee ancestor found east of the rift (*A. ramidus* according to ref. 8) be consistent with such a model? It is vastly more likely that *Ardipithecus* is not a member of the chimpanzee clade, because of the many derived characters it shares with later hominids¹¹. It is also clear that more information will be needed to resolve the role of *Orrorin* in hominoid phylogeny.

Likewise, the phylogenetic and taxonomic status of the Middle Awash fossils described here will require review as hypodigms increase. They appear to represent a hominid situated temporally and anatomically close to the last common ancestor of chimpanzees and humans. These Late Miocene fossils are followed temporally in the Middle Awash by a 5-Myr succession of increasingly derived hominid taxa, including *Ardipithecus ramidus*, *Australopithecus afarensis*, *Australopithecus garhi* and species of *Homo*. □

Methods

Description

Primates Linnaeus, 1758

Anthropoidea Mivart, 1864

Hominoidea Gray, 1825

Hominidae Gray, 1825

Ardipithecus White, Suwa and Asfaw, 1995

Ardipithecus ramidus (White, Suwa and Asfaw, 1994)

Ardipithecus ramidus kadabba subsp. nov.

Etymology. The subspecific name, kadabba, is taken from the Afar language. It means basal family ancestor.

Holotype. ALA-VP-2/10 (Fig. 1) is a right mandibular corpus with M₃, left I₂, C, P₄, M₂ and M₃ root fragment. Holotype and referred material are housed at the National Museum of Ethiopia, Addis Ababa. Holotype from Alayla Vertebrate Paleontology Locality Two (ALA-VP 2); differentially corrected GPS coordinates 10° 16.483' N and 40° 15.313' E; elevation 690 m.

Referred material. ALA-VP-2/11 (intermediate hand phalanx); ALA-VP-2/101 (left humerus and ulna); ASK-VP-3/78 (distal humerus); ASK-VP-3/160 (left P₃); DID-VP-1/80

(proximal hand phalanx fragment); STD-VP-2/61 (right lower canine); STD-VP-2/62 (LM²); STD-VP-2/63 (LM¹); STD-VP-2/893 (left clavicle fragment); AME-VP-1/71 (proximal foot phalanx).

Localities. Saitune Dora (STD-VP-2), Alayla (ALA-VP-2), Asa Koma (ASK-VP-3) and Digiba Dora (DID-VP-1) are all located along the western margin of the Middle Awash study area in the Afar depression of Ethiopia. Amba East (AME-VP-1) is in the CAC^{7,10}.

Horizons. The four western-margin hominid localities discussed here are within the Asa Koma Member of the Adu Asa Formation and bracketed by an overlying basaltic flow dated to 5.54 ± 0.17 Myr and an underlying basaltic tuff dated to 5.77 ± 0.08 Myr⁷. The Amba hominid is from the Kuserale Member of the Sagantole Formation of the CAC and is bracketed to $5.2\text{--}5.6$ Myr¹⁰.

Diagnosis. On the limited available evidence, a subspecies of *Ardipithecus* distinguished from *Aramis A. ramidus* (*A. ramidus ramidus*) by sharp M₃ lingual cusps that retain their prominence even in extreme crown wear; squared distal outline to M³ with four distinct cusps; shallow mesial fovea on P³; tendency for less relief on the mesiolingual crown face of the lower canines (one of two specimens); mesiolingually-to-distobuccally compressed lower canines.

Ardipithecus ramidus kadabba is distinguished from fossil and extant apes in its tendency toward incisiform lower canines, comparable to the condition of *Aramis A. ramidus*, with a developed distal tubercle and variants with high mesial crown shoulder placement and some expression of the mesial marginal ridge.

Received 19 February; accepted 15 May 2001.

1. Ruvolo, M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol.* **14**, 248–265 (1997).
2. Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl Acad. Sci. USA* **92**, 532–536 (1995).
3. Chen, F.-C. & Li, W.-H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
4. Hill, A. in *Integrative Paths to the Past: Paleoanthropological Advances in Honor of F. Clark Howell* (eds R. S. Corruccini and R. L. Ciochon) 123–145 (Prentice Hall, Englewood Cliffs, 1994).
5. McDougall, I. & Feibel, C. Numerical age control for the Miocene–Pliocene succession at Lothagam, a hominoid-bearing sequence in the northern Kenya Rift. *J. Geol. Soc. Lond.* **156**, 731–745 (1999).
6. Leakey, M. G. *et al.* A record of faunal change in the late Miocene of East Africa. *J. Vert. Paleontol.* **16**, 556–570 (1996).
7. WoldeGabriel, G. *et al.* Geology and palaeontology of the Late Miocene Middle Awash valley, Afar rift, Ethiopia. *Nature* **412**, 175–178 (2001).
8. Senut, B. *et al.* First hominid from the Miocene (Lukeino Formation, Kenya). *C.R. Acad. Sci. Ser. IIa* **332**, 137–144 (2001).
9. Pickford, M. & Senut, B. The geological and faunal context of Late Miocene hominid remains from Lukeino, Kenya. *C.R. Acad. Sci. Ser. IIa* **332**, 145–152 (2001).
10. Renne, P. R., WoldeGabriel, G., Hart, W. K., Heiken, G. & White, T. D. Chronostratigraphy of the Miocene–Pliocene Sagantole Formation, Middle Awash Valley, Afar rift, Ethiopia. *Geol. Soc. Am. Bull.* **111**, 869–885 (1999).
11. White, T. D., Suwa, G., & Asfaw, B. *Australopithecus ramidus*, a new species of hominid from Aramis, Ethiopia. *Nature* **371**, 306–312 (1994).
12. Latimer, B., Lovejoy, C. O., Johanson, D. C. & Coppens, Y. Hominid tarsal, metatarsal, and phalangeal bones recovered from the Hadar Formation: 1974–1977 collections. *Am. J. Phys. Anthropol.* **57**, 701–719 (1982).
13. Latimer, B. & Lovejoy, C. O. Metatarsophalangeal joints of *Australopithecus afarensis*. *Am. J. Phys. Anthropol.* **83**, 13–23 (1990).
14. Day, M. H. Femoral fragment of a robust australopithecine from Olduvai Gorge, Tanzania. *Nature* **47**, 230–233 (1969).
15. Ohman, J. C., Krochta, T. J., Lovejoy, C. O., Mensforth, R. P. & Latimer, B. Cortical bone distribution in the femoral neck of hominoids: implications for the locomotion of *Australopithecus afarensis*. *Am. J. Phys. Anthropol.* **104**, 117–131 (1997).

Acknowledgements

The National Science Foundation, the Wenner-Gren Foundation, and the University of California at Berkeley provided funding. The Authority for Research and Conservation of Cultural Heritage of the Ministry of Information and Culture granted field permits, and the National Museum of Ethiopia granted access to the Paleoanthropological Laboratory before 13 February 2001. The success of this research is largely owed to the members of the Middle Awash research project and the Afar people. This contribution is dedicated to our late friend Neina Tahiru, who suggested the name '*kadabba*', and to whom I give special thanks. G. WoldeGabriel played a major role in discovering fossiliferous hominid-bearing localities along the western margin of the Middle Awash study area and also studied the geology of the region. I thank T. White, B. Latimer, K. Geleta, H. Gilbert, D. DeGusta, L. Hlusko, E. Gülec, C. Pehlevan, B. Asfaw, M. Black, G. Suwa, S. Yosef, A. Amzaye, M. Asnake and H. Saegusa for their participation in survey and excavations. Sheikh Ebrahim and Shiekh Oumer helped coordinate the Afar labour force. I thank B. Latimer, O. Lovejoy and S. Simpson for their assistance during the comparative studies conducted at the Cleveland Museum of Natural History. O. Lovejoy, T. White and G. Suwa provided insights and comments. Advice, support and encouragement to conduct research along the western margin of the Middle Awash study area were extended by T. White and B. Asfaw.

Correspondence and requests for materials should be addressed to Y.H.-S. (e-mail: ethio@uclink4.berkeley.edu).

Rapid and recent origin of species richness in the Cape flora of South Africa

James E. Richardson*†, Frans M. Weitz‡§, Michael F. Fay*, Quentin C. B. Cronk¶, H. Peter Linder†§, G. Reeves* & Mark W. Chase*

* Jodrell Laboratory, Royal Botanic Gardens, Richmond, Surrey TW9 3DS, UK

‡ Department of Botany, University of Western Cape, Bellville 7535, Cape Province, South Africa

§ Bolus Herbarium, Botany Department, University of Cape Town, Rondebosch 7700, South Africa

¶ Institute of Cell and Molecular Biology, University of Edinburgh, Darwin Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, UK

¶ Royal Botanic Garden, 20A Inverleith Row, Edinburgh EH3 5LR, UK

The Cape flora of South Africa grows in a continental area with many diverse and endemic species^{1–4}. We need to understand the evolutionary origins and ages of such 'hotspots' to conserve them effectively⁵. In volcanic islands the timing of diversification can be precisely measured with potassium–argon dating. In contrast, the history of these continental species is based upon an incomplete fossil record and relatively imprecise isotopic palaeotemperature signatures. Here we use molecular phylogenetics and precise dating of two island species within the same clade as the continental taxa to show recent speciation in a species-rich genus characteristic of the Cape flora. The results indicate that diversification began approximately 7–8 Myr ago, coincident with extensive aridification caused by changes in ocean currents. The recent origin of endemic species diversity in the Cape flora shows that large continental bursts of speciation can occur rapidly over timescales comparable to those previously associated with oceanic island radiations^{6,7}.

Phyllica is a genus of the buckthorn family (Rhamnaceae) with approximately 150 species, most of which occur in the Cape province, but it is also disjunctly distributed through other parts of southern Africa. Species of *Phyllica* also occur on several oceanic, volcanic islands, such as St Helena⁸, which enabled us to calibrate a molecular clock both within and outside this radiation. Two critical points of calibration were known, and these two bracketing dates provide mutual checks for the accuracy of the dating: (1) dispersal of one species from Mauritius to the volcanic island of Réunion, which is known to be two million years old, provided us with a calibration point for a relationship within *Phyllica*; and (2) *Nesiota*, a closely related, endemic genus on St Helena (known to be 14.3 million years old), served as an external calibration point. We used DNA sequence data, from both the plastid and nuclear genome, to reconstruct the biogeographical history of *Phyllica* and related genera, demonstrating that rapid species diversification in *Phyllica* took place from 7–8 Myr ago, with most species appearing even more recently.

Aridification in the Cape region has been attributed to the separation of Antarctica from South America, which allowed a cold circum-Antarctic (Benguela) current to develop around 11–14 Myr ago. Such aridification may have been an important factor in initiating the transformation of the Miocene (26 Myr ago) sub-tropical forest to the fynbos vegetation of today^{9–12}. The sparse fossil record of the Cape gives evidence of changes in the ecological dominance of genera typical of the fynbos, but it cannot provide direct evidence of increasing species richness. Previously there has

† Present address: Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95064, USA (J.E.R.); Institute for Systematic Botany, Zollikerstrasse 107, CH8008, Zurich, Switzerland (H.P.L.); Kirstenbosch Research Centre, National Botanical Institute, Rhodes Drive, Newlands 7735, Cape Town, South Africa (G.R.).

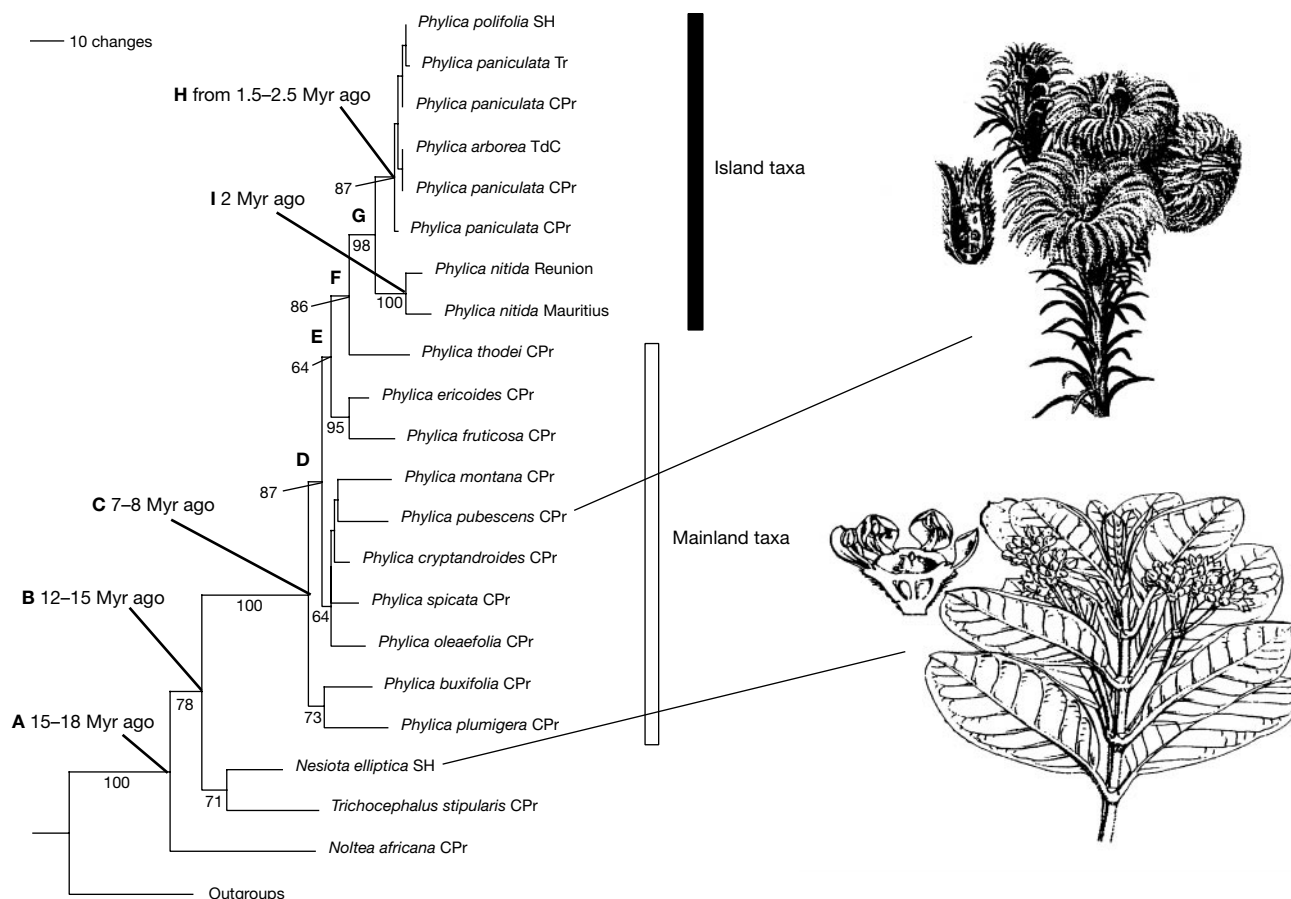


Figure 1 One of the six phylogenetic trees. The tree was randomly selected from the parsimony analysis of the combined nuclear and plastid data (tree length 916 steps with bootstrap percentages indicated below the branches. Branch lengths are proportional to the changes in the internal transcribed spacers nuclear ribosomal DNA data. TdC, Tristan da Cunha; SH, St Helena; NA, New Amsterdam; CPr, Cape Province; Tr, Transvaal. Outgroup taxa have been removed to show more details within the *Phylica* portion of the

tree. Species illustrated to the right (flowering branch and a cross-section of a flower) are *Phylica pubescens*, which is typical of the highly modified species for the Cape of South Africa, and *Nesiota elliptica* (extinct in the wild), which has typical unmodified morphology for Rhamnaceae (broad leaves and unspecialized flowers with a generalized pollination syndrome).

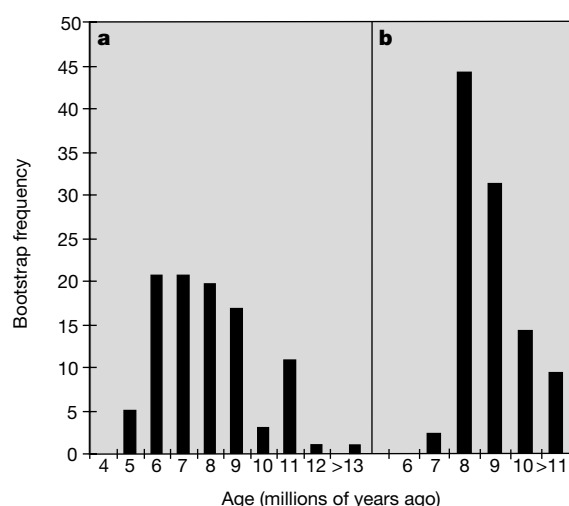


Figure 2 Histograms of bootstrap results using two different nodes to calibrate the molecular clock for *Phylica*. **a**, Bootstrap frequencies for the *P. nitida* node (node I, Fig. 1) as the calibration point, mean age 7.4 Myr, standard deviation 0.18. **b**, Bootstrap frequencies for the *Nesiota/Trichocephalus* and *Phylica* node (node B, Fig. 1) as the calibration point, mean age 8.45 Myr, standard deviation 0.13. The deeper node at which sequence divergence is greater provides a more consistent age estimate with the bootstrapping procedure, but both calibration points are in good agreement about the age of the *Phylica* radiation.

been no conclusive way to connect recent aridification to an increase in diversity, but our results using a molecular clock now give support to previous ideas based on interpretation of the fragmentary fossil and climatic data. Although speciation in the Cape was primarily driven by aridification¹³, it was undoubtedly also aided by a number of other factors, including a high diversity of soil types, complex physiography with each mountain peak having a distinct climate¹⁴, reproductive isolation of populations in transient fire-created niches, and pollinator shifts¹⁵. In addition there have been regular climatic fluctuations in the form of glacial cycles during the Pleistocene causing periodic fragmentation of ranges, leading to isolation of small populations. The extensive diversification of *Phylica* can be compared with other examples of recent, explosive speciation^{6,7,16,17} and accurately demonstrates that these processes can take place over similar timescales on continents as well as on oceanic island archipelagos where these radiations have been previously well studied and documented. We need to understand the forces behind the production of such biotic diversity if conservation of what remains is to be successful, especially in the light of potential climate change in the future.

Methods

Gene sequencing

Sequences of both nuclear ribosomal (internal transcribed spacers, ITS) and plastid DNA (*trnL-F* intron and intergenic spacer) were produced for island species and representatives of the major morphological and geographical groups of *Phylica* in the Cape (each of which

consists of about 10 species) and taxa that were determined as close relatives in a complete analysis of Rhamnaceae¹⁸. The data matrix containing the aligned sequences is available on request from the J.E.R. All sequences have been submitted to GenBank (accession numbers AJ327603–327621, AJ328801–328835, AJ225798, AJ225803, AJ390350, AJ390352–390353, AJ390356–390358, AJ390361, AJ390363). The species included here are a subset of those in a larger matrix and represent each of the lineages identified (F.M.W., J.E.R., M.F.F., Q.C.B.C., M.W.C. and H.P.L., unpublished work).

Phylogenetic analysis

Data were analysed using the parsimony algorithm of the software package PAUP* version 4.0d64 for Macintosh¹⁹. Tree searches were conducted under the equal and unordered weights criterion (Fitch parsimony)²⁰ with 1,000 random sequence additions and tree bisection reconnection branch swapping. One thousand replicates of the bootstrap²¹ were used to estimate tree stability. Parsimony analysis of both separate and combined data sets produced trees with the same topology and high bootstrap percentages for all lineages that affected the dating. The combined analysis produced only six shortest trees, among which differences involved variation in portions of the trees not important to the conclusion reached here (for example, within the core clades of *Phylica*).

Molecular clock calibration

Rate heterogeneity among lineages was evaluated for the ITS data set using the likelihood ratio test²², which compares log likelihoods of both constrained and unconstrained hypotheses. We used a three-parameter, maximum-likelihood model with transition/transversion ratio and gamma distribution of rate variation among sites estimated from the ITS sequence region. The molecular clock was rejected because the constrained and unconstrained analyses were significantly different (4,905.3 versus 4,827.3; $P < 0.005$), so Sanderson's method of nonparametric rate smoothing (NPRS)²³ was applied to produce an ultrametric tree using TreeEdit version 1.0 alpha 4-61 (ref. 24). We used the ITS data optimized on the combined tree (shown in Fig. 1) to calibrate the tree at two nodes in absolute time: (1) a date of 2 Myr ago was used to define the split (node I, Fig. 1) between the Mauritian and Réunion populations of *Phylica nitida* (both shown with DNA fingerprinting by AFLP analysis to be monophyletic; J.E.R., M.F.F., Q.C.B.C. and M.W.C., manuscript in preparation); (2) a date of 14.3 Myr ago was used to define the split (node B; Fig. 1) of the sister clade of *Phylica* and that of *Phylica* itself. We also determined the confidence intervals for these dates by keeping the tree fixed and bootstrapping the data set 100 times, each time estimating the divergence time on this single fixed tree; this allowed us to construct a histogram of the possible divergence times (Fig. 2). The length of the NPRS branch of *P. nitida* on Réunion enabled us to infer dates for the other nodes within the phylogenetic tree for *Phylica* and related taxa. *Noltea* (node A) split from the *Phylica* lineage 15–18 Myr ago. Node B, that of *Nesiota* and *Trichocephalus* relative to *Phylica*, dates to 12–15 Myr ago, which accords well with the putative dispersal of the ancestor of *Nesiota* to St Helena, the origin of which is dated at 14.3 Myr ago. Node C, at 7–8 Myr ago, which is the one marking the onset of species diversification of *Phylica* in the Cape, parallels the point at which rapid aridification in the Cape began. Node H occurred at 1.5–2.5 Myr ago, which accords well with the age of the youngest islands on which *P. arborea* is endemic in the southern oceans. Dates of nodes calibrated from the most parsimonious tree (uncorrected branch lengths) provide older age estimates, but these fall within those calculated from the bootstrap of the NPRS tree. NPRS is known to be inaccurate if sequence divergence is low (which is true for the first calibration point used above, but not for the second one deeper in the tree), so we also estimated dates without NPRS (based solely on the maximum-likelihood estimate of branch lengths). These results all fall within the confidence intervals defined by the bootstrapping procedure. Our conclusions about the timing of the *Phylica* radiation and its connection to the most recent period of aridification are thus supported regardless of the method of estimation used; we prefer the dates from NPRS because we know that rate heterogeneity is present in these ITS data, and the dates presented above and in Fig. 1 are the NPRS dates.

Received 17 January; accepted 29 May 2001.

- Good, R. *The Geography of Flowering Plants* Ch. 10 190–191 (Longman, London, 1974).
- Goldblatt, P. An analysis of the flora of southern Africa: its characteristics, relationships and origins. *Ann. Missouri Bot. Gard.* **65**, 369–436 (1978).
- Takhtajan, A. *Floristic Regions of the World* Ch. 4 263–267 (Univ. California Press, Berkeley, 1986).
- Linder, H. P., Meadows, M. E. & Cowling, R. M. in *The Ecology of Fynbos: Nutrients, Fire and Diversity* (ed. Cowling, R. M.) 113–134 (Oxford Univ. Press, Cape Town, 1992).
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
- Baldwin, B. G. & Robichaux, R. H. in *Hawaiian Biogeography: Evolution on a Hotspot Archipelago* (eds Wagner, W. L. & Funk, V. A.) 257–287 (Smithsonian Institution Press, Washington DC, 1995).
- Givnish, T. J., Sytsma, K. J., Smith, J. F. & Hahn, W. J. in *Hawaiian Biogeography: Evolution on a Hotspot Archipelago* (eds Wagner, W. L. & Funk, V. A.) 288–337 (Smithsonian Institution Press, Washington DC, 1995).
- Cronk, Q. C. B. *The Endemic Flora of St Helena* 23 (Anthony Nelson, Oswestry, 2000).
- Axelrod, D. I. & Raven, P. H. in *Biology and Ecology of Southern Africa* (ed. Werger, M. J. A.) 77–130 (Junk, The Hague, 1978).
- Coetzee, J. A. in *Antarctic Glacial History and World Palaeoenvironments* (ed. Van Zinderen Bakker, E. M.) 115–127 (Balkema, Rotterdam, 1978).
- Coetzee, J. A. Intimations on the Tertiary vegetation of southern Africa. *Bothalia* **14**, 345–354 (1983).
- Kennet, J. P. Palaeoceanographic and biogeographic evolution of the southern ocean during the Cenozoic, and Cenozoic microfossil datums. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **31**, 123–152 (1980).
- Linder, H. P. & Mann, D. M. The phylogeny and biogeography of *Thamnochortus* (Restionaceae). *Bot. J. Linn. Soc.* **128**, 319–357 (1998).

- Linder, H. P. in *Species and Speciation* (ed. Vrba, E. S.) 53–57 (Transvaal Museum Monograph 4 Pretoria, 1985).
- Johnson, S. D. Pollination, adaptation and speciation models in the Cape flora of South Africa. *Taxon* **45**, 59–66 (1995).
- Hodges, S. A. & Arnold, M. L. Columbines: a geographically widespread species flock. *Proc. Natl Acad. Sci. USA* **92**, 5129–5132 (1994).
- Wojciechowski, M. F., Sanderson, M. J. & Hu, J.-M. Evidence on the monophyly of *Astragalus* (Fabaceae) and its major subgroups based on nuclear ribosomal QNA ITS and chloroplast DNA *trnL* intron data. *Syst. Bot.* **24**, 409–437 (1999).
- Richardson, J. E., Fay, M. F., Cronk, Q. C. B., Bowman, D. & Chase, M. W. A molecular analysis of *Rhamnaceae* using plastid *rbcl* and *trnL-F* sequences. *Am. J. Bot.* **87**, 1309–1324 (2000).
- Swofford, D. L. PAUP* 4.0b2: Phylogenetic Analysis Using Parsimony. (Sinauer Associates, Sunderland, Massachusetts, 1998).
- Fitch, W. M. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.* **20**, 406–416 (1971).
- Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
- Felsenstein, J. Evolutionary trees from DNA-sequences—a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
- Sanderson, M. J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1232 (1997).
- Rambaut, A. & Charleston, M. TreeEdit version 1.0 alpha 4-61 (<http://evolve.zoo.ox.ac.uk/software/TreeEdit/TreeEdit.html>) (2000).

Acknowledgements

We are grateful to P. Crane, M. Sanderson and the Tropical Biology Group at the Royal Botanic Garden, Edinburgh, for critical comments and discussion. We thank A. de Bruijn and J. Joseph for technical support. The work was funded by a studentship to J.E.R. from the Royal Botanic Gardens, Kew, which also made possible a four-month visit by F.M.W. to Kew to collect additional data. We also thank collectors of plant material: Y. Mungroo, C. Thébaud, M. van der Bank and R. Cairns-Wicks.

Correspondence and requests for material should be addressed to J.E.R. (e-mail: jamesr@darwin.ucsc.edu).

Predators increase the risk of catastrophic extinction of prey populations

Thomas W. Schoener*, David A. Spiller* & Jonathan B. Losos†

* Section of Evolution and Ecology, University of California, Davis, California 95616, USA

† Department of Biology, Washington University, St Louis, Missouri 63130, USA

There has been considerable research on both top-down effects^{1,2} and on disturbances^{3–5} in ecological communities; however, the interaction between the two, when the disturbance is catastrophic, has rarely been examined⁶. Predators may increase the probability of prey extinction resulting from a catastrophic disturbance both by reducing prey population size^{7,8} and by changing ecological traits of prey individuals such as habitat characteristics^{8,9} in a way that increases the vulnerability of prey species to extinction. We show that a major hurricane in the Bahamas led to the extinction of lizard populations on most islands onto which a predator had been experimentally introduced, whereas no populations became extinct on control islands. Before the hurricane, the predator had reduced prey populations to about half of those on control islands. Two months after the hurricane, we found only recently hatched individuals—apparently lizards survived the inundating storm surge only as eggs. On predator-introduction islands, those hatchling populations were a smaller fraction of pre-hurricane populations than on control islands. Egg survival allowed rapid recovery of prey populations to pre-hurricane levels on all control islands but on only a third of predator-introduction islands—the other two-thirds lost their prey populations. Thus climatic disturbance compounded by predation brought prey populations to extinction.

The experiment, which was punctuated by the hurricane, began in April 1997 when the large predatory lizard *Leiocephalus carinatus* was introduced onto five islands that already had populations of a smaller lizard, *Anolis sagrei*. Six other islands with *A. sagrei* populations served as controls; we also monitored an island that *L. carinatus* had invaded naturally shortly before the experiment began; this island is counted with the islands that were experimentally invaded by *Leiocephalus carinatus* below. *Leiocephalus carinatus* colonist lizards were collected from the mainland of Great Abaco, the Bahamas, using sites ranging from the parts of the mainland that are very close to the study islands to areas up to 12 km away.

The larger species of lizard, which is known to consume the smaller *A. sagrei*¹⁰, had immediate and dramatic effects on the invaded species. Although introduction and control islands averaged similar numbers of *A. sagrei* at the beginning, the population size on islands where *L. carinatus* was introduced dropped markedly, levelling off at about half the control value (Fig. 1a). The smaller species shifted its habitat upwards in the vegetation, eventually occupying the higher and thinner branches (Fig. 1b), which the larger species mostly avoided. Pre-hurricane results of this experiment are being reported elsewhere¹¹.

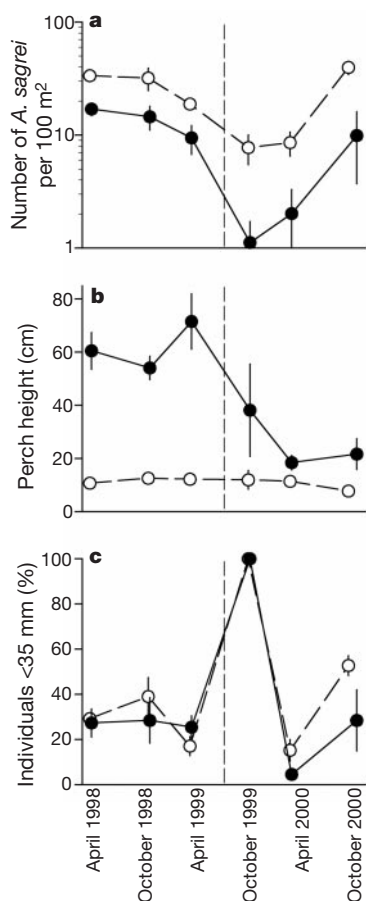


Figure 1 Properties of populations of the lizard *Anolis sagrei* before and after hurricane Floyd. Vertical dotted line marks the occurrence of the hurricane. Open circles, *Leiocephalus carinatus* absent; closed circles, *L. carinatus* introduced. **a**, Density of *A. sagrei* on the two categories of study islands (with and without the larger predatory lizard). **b**, Perch height of *A. sagrei*. **c**, The fraction of very small *A. sagrei* individuals, that is, those likely to have hatched from eggs within the past two months. In another study¹¹, we showed that *L. carinatus* significantly reduced the smallest (<29 mm) lizards; here, using a size of <35 mm, a similar difference is apparent at most dates except the one immediately after the hurricane, when all individuals in both treatments are small. We note that the first year of the pre-hurricane portion of the experiment is not shown.

On 14 September 1999, Floyd, a category-IV hurricane with maximum sustained winds of around 250 km h⁻¹ and an approximately 3-m storm surge passed directly over the study site. Two months later, 13–19 November, we saw 65 different individuals of *A. sagrei* on all 12 study islands combined; none was larger than 34 mm and only one was larger than 32 mm, a dramatic contrast with a typical year's distribution in size of individuals (Fig. 1c). Thus the post-hurricane lizard populations consisted entirely of juveniles! Figure 2 illustrates the huge change that the hurricane wrought upon the size-frequency distribution of *A. sagrei*. (That this post-hurricane cohort was growing up in the total absence of adults is not necessarily deleterious, as *A. sagrei* has no parental care and indeed shows occasional cannibalism¹².) Although only three individuals of the large species, *L. carinatus*, were seen (one on each of three islands), these were also relatively small. We conclude that the lizards must have survived the hurricane in the egg stage. This discovery implies that lizard eggs can survive immersion in salt water for some time—in this case, up to 6 h based on estimates of storm-surge duration by locals. We are now testing this experimentally and have found that younger (≤10 days old) eggs immersed in sea water for 3 or 6 h do not differ in viability from controls; experiments on older eggs are underway.

The hurricane left the islands divided into one of two discrete states with respect to *A. sagrei*. Each island had either a substantial number of juvenile *A. sagrei*, possibly enough to 'seed' a new population, or it had zero individuals recorded during at least one time (Fig. 3). Populations of *A. sagrei* on all eight islands of the first state re-attained pre-hurricane levels in about 14 months. One of the four islands of the second state had never contained *A. sagrei* populations (N1), whereas two others (A18, X10) had one, and the fourth (Z3) had two *A. sagrei* at the last census; the individual on X10 and at least one on Z3 were almost certainly natural colonists—those islands are very close to a mainland source.

Which of the two states an island occupied was not random with respect to the presence of the large predator, *L. carinatus*. All six

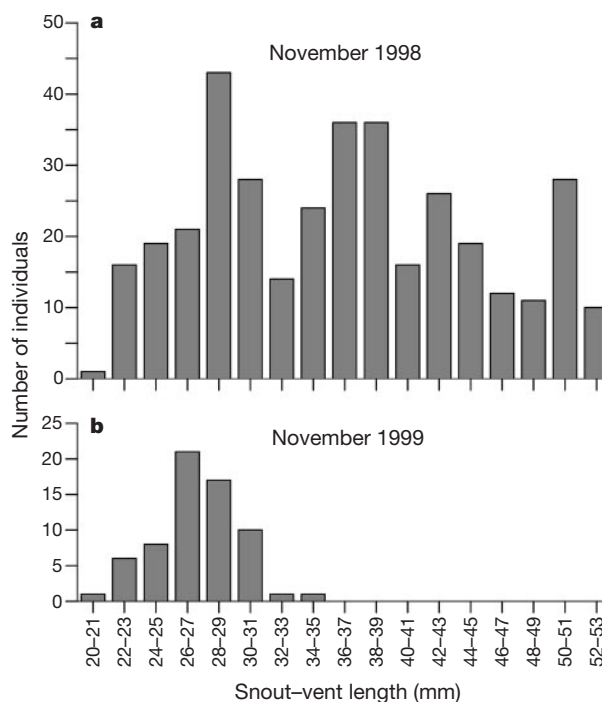


Figure 2 The abundance distribution of *A. sagrei* on all islands combined before (**a**) and after (**b**) hurricane Floyd. The sizes are approximate, that is, they were estimated visually. $N = 360$ (**a**) and 65 (**b**).

islands without the large predator showed the pattern of complete recovery, whereas four of the islands which had *L. carinatus* populations before the hurricane showed the pattern of extinction or extreme population reduction. The overall pattern, four of six populations becoming extinct on predator-introduction islands versus zero of six on control islands, is significant with a Fisher exact test (one-tailed $P = 0.030$).

Although *L. carinatus* had diminished populations of *A. sagrei* to about half the control average by the time of the census immediately preceding the hurricane, population-size rank itself was not a significant predictor of extinction versus survival (Wilcoxon–Mann–Whitney test: one-tailed $P = 0.107$, $n = 12$). Indeed, the presence of the large predator increased the proportional amount by which *A. sagrei* populations were lowered after the hurricane: the percentage drop from the census before the hurricane to the census taken two months afterwards averaged 61.0 for islands without *L. carinatus* and 85.7 for islands with that predator (Median test for difference in percentage decline: one-tailed $P = 0.014$). This unexpected result, although it is not independent of the result for predator introduction versus extinction, provides insight into how predation acts synergistically with disturbance to cause extinction.

To explain it we offer three hypotheses: first, immediate post-hurricane predation by the *L. carinatus* that survived as eggs on some of the islands may have directly reduced *A. sagrei* hatchling numbers; thus the continuing presence of the predator is important. This hypothesis is supported by the fact that the two predator-introduction islands on which *A. sagrei* recovered fully (Z2, Z4) had no *L. carinatus* found after the hurricane; on only one of the other four predator-introduction islands were no *L. carinatus* found. Evidence against this first hypothesis is the size of the *L. carinatus* that survived the hurricane, which were hatchlings themselves and may thus have been less able to feed on prey as large as other lizards.

Second, because of the habitat shift of *A. sagrei*, eggs may have been deposited in less secure places, such as high in trees rather than in rock holes or other more protected sites. This is supported by the extreme habitat shift forced upon *A. sagrei* by the large predator: the higher, thinner perches which *A. sagrei* favoured after the introduction of *L. carinatus* (Fig. 1) probably do not have as many secure sites for depositing eggs. For example, treetops would be potentially more vulnerable to the various forces of the storm: wind, driving rain, high waves and the storm surge itself.

Third, the rate of egg production per female may have been reduced by the large predator. This may also be related to the habitat shift: the unsuitable new surroundings may result in reduced food intake and mating frequency in *A. sagrei* thereby lowering reproductive rate. Although we have no data on mating, food intake should be reflected by body condition (weight relative to length); this was significantly reduced on *L. carinatus* islands in adult males but not in females¹¹. Small population-size effects could themselves also contribute to lower reproductive rates; however, percentage reduction was not correlated with population size before the hurricane (Spearman rank correlation = -0.28 , $P > 0.25$, $n = 12$).

Ecologists have long debated the extent to which communities are shaped by biotic versus physical factors^{5,13–18}. For example, do the distribution and abundance of species depend more upon the competitors and/or predators present or upon the vicissitudes of climate? The degree to which one or the other is important to some extent determines the predictability of systems, that is, species interactions often have predictable results whereas climatic events and dispersal during recovery from these events are more stochastic. In the present situation, a highly predictable biological interaction between the larger and smaller lizard species reduced the smaller to a population size, and perhaps reproductive state, whereby an unexpected climatic catastrophe caused its extinction on several islands. The timing of the extinctions was not predictable except in the very

short term, but which islands suffered the extinctions was very predictable. Thus, a compounding of (biotic) predation and (physical) cataclysm basically wiped out a species on four islands where it was a prey species rather than being at the top of its food web, a position it enjoyed on all the islands where it survived. The fact that *A. sagrei* is able to consume a very wide variety of prey allows it to maintain very high chronic densities on islands without predators^{19,20} and to recover on those islands as well. The larger predator, *L. carinatus*, also has a very general diet¹⁰, yet because of its size, is unable to exist at nearly as high densities: before the hurricane, *L. carinatus* (even with artificial supplementation) maintained numbers about an order of magnitude lower than those of *A. sagrei*—6.9 versus 54.4 individuals per island. Indeed, *L. carinatus* became extinct on four of its six islands by November 2000.

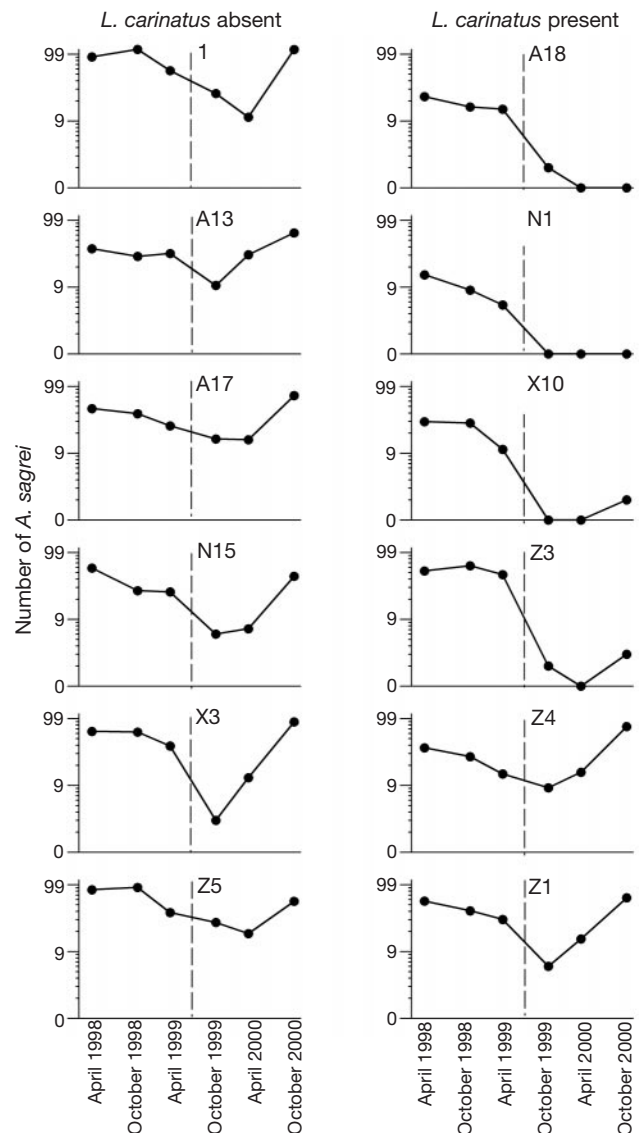


Figure 3 Mortality and recovery or extinction of the lizard *A. sagrei* from hurricane Floyd. Population size is plotted for each census date. Vertical dotted line marks the occurrence of the hurricane. The left column shows control islands, without the large predatory lizard *L. carinatus*. The right column shows islands with the large predator introduced. Islands fall into two patterns: (1) those completely recovering, all of which were permanently without the large predator, plus Z4 and Z1 (both of which lost the large predator as a result of the hurricane); versus (2) islands with extinction or extreme population reduction (N1, A18, X10 and Z3). Three of the latter islands (all but A18) had the large predator at least some of the time after the hurricane.

How widespread is the phenomenology we have been able to demonstrate? The reduction and even extinction of populations by a combination of biotic and physical mortality factors should be quite common. Indeed, we suggested this general mechanism for the extinction of spider species on smaller islands of the same archipelago, but we had no direct evidence²⁰. Physical disturbance that is catastrophic, such as that resulting from hurricanes, is itself less common of course, so co-action with a biotic factor such as predation would perforce be even less common; the spider example just given involves more frequent, less catastrophic storms. The unanticipated nature of hurricanes makes even rarer the observations necessary to record their effects; a measure of serendipity is required²¹. But this may change: the gradual increase in sea surface temperatures, perhaps related to climate change^{22,23}, appears to be contributing to the increase in hurricane frequency. In fact, five of the last six years have had a greater than average number of Category III or more severe hurricanes (see <http://weather.unisys.com/hurricane/atlantic/index.html>); one must go back to 1969 to accumulate another five years that are above average. Thus, the direct effects of climatic catastrophes as well as interactive effects with biotic factors, some producing extinctions, may be on the rise. Conservation biology might profit from consideration of both kinds of effects in analysing extinction risk. □

Methods

Study site and subject species

The study islands ranged in vegetated area from 137 to 270 m² ($\bar{x}$ = 191.0). All were covered with fairly closely spaced shrubs and trees seldom exceeding 2 m. All were located in an approximately 7 × 2 km area that encompassed several protected 'creek' waterways fanning out from Buckaroon bay and Snake cay, Great Abaco, the Bahamas.

The manipulated species was the lizard *Leiocephalus carinatus armouri*, a subspecies native to the Little Bahama bank²⁴; it is one of the larger forms of the species¹⁴ (maximum snout–vent length: male, 107 mm; female, 97 mm). The subject lizard species *A. sagrei* is substantially smaller (maximum snout–vent length: male, 57 mm; female, 44 mm). *Leiocephalus carinatus* prefers substantially lower perches than does *A. sagrei*²⁵, and it is most frequently found on the ground. Although it is thus more arboreal than *L. carinatus*, *A. sagrei* in more diverse lizard communities perches lower than other *Anolis* lizards²⁶.

Experimental protocol

To select islands for introductions, we first (in 1996) stratified the 12 islands with populations of *A. sagrei* into six pairs according to area, vegetation structure and number of *A. sagrei*. One of the islands was naturally colonized in the ensuing year (see above); we randomly chose one of each of the remaining five pairs of islands for the introduction of *L. carinatus*. Throughout the experiment, we attempted to maintain each *L. carinatus* population at five or more individuals by further introductions when necessary. After 1.5 yr, the populations stabilized on all islands and we did not have to add more.

Lizard censuses

Lizard numbers were estimated using the multiway-contingency-table procedure^{27,28}; three censuses are done on each of three different days, during which lizards are marked with a census-specific colour of water-soluble latex paint administered with Indico long-distance spraying devices. We used a multivariate-contingency-table method to fit the data to each of eight models, distinguished by which interactions (associations between the three censuses), if any, are included. The simplest of the models fitting the data adequately was selected and the estimate provided by that model was used. A Pascal program doing the fits was kindly provided by J. Roughgarden. Occasionally the program ran no model on the data; then we used the average of the possible pairwise Lincoln estimates (further details in ref. 11). We performed Fisher exact and median tests with SAS software; and used the Wilcoxon–Mann–Whitney test from ref. 29.

Received 2 March; accepted 10 May 2001.

1. Pace, M. L., Cole, J. J., Carpenter, S. R. & Kitchell, J. F. Trophic cascades revealed in diverse ecosystems. *Trends Ecol. Evol.* **14**, 483–488 (1999).
2. Schmitz, O. J., Hamback, P. A. & Beckerman, A. P. Trophic cascades in terrestrial systems: a review of the effects of carnivore removals on plants. *Am. Nat.* **155**, 141–153 (2000).
3. Sousa, W. P. The role of disturbance in natural communities. *Annu. Rev. Ecol. Syst.* **15**, 353–392 (1984).
4. Pickett, S. T. A. & White, P. S. *The Ecology of Natural Disturbance and Patch Dynamics* (Academic, New York, 1985).
5. Dunson, W. A. & Travis, J. The role of abiotic factors in community organization. *Am. Nat.* **101**, 97–107 (1991).
6. Power, M. E., Parker, M. S. & Wootton, J. T. in *Food Webs* (eds Polis, G. A. & Winemiller, K. O.) 286–297 (Chapman and Hall, London/New York, 1996).
7. Sih, A., Crowley, P., McPeck, M., Petranka, J. & Strohmeier, K. Predation, competition, and prey communities: a review of field experiments. *Annu. Rev. Ecol. Syst.* **16**, 269–311 (1985).

8. Abrams, P. A. Implications of dynamically variable traits for identifying, classifying, and measuring indirect effects in ecological communities. *Am. Nat.* **146**, 112–134 (1996).
9. Schmitz, O. J., Beckerman, A. P. & O'Brien, K. M. Behaviourally-mediated trophic cascades: the effects of predation risk on food web interactions. *Ecology* **78**, 1388–1399 (1997).
10. Schoener, T. W., Slade, J. B. & Stinson, C. H. Diet and sexual dimorphism in the very catholic lizard genus *Leiocephalus* of the Bahamas. *Oecologia* **53**, 160–169 (1982).
11. Schoener, T. W., Spiller, D. A. & Losos, J. B. Predation upon a devastating predator: can the latter's food-web effects be reversed? *Ecol. Monogr.* (submitted).
12. Gerber, S. P. & Echternacht, A. C. Evidence for asymmetrical intraguild predation between native and introduced *Anolis* lizards. *Oecologia* **124**, 599–607 (2000).
13. Andrewartha, H. G. & Birch, L. C. *Distribution and Abundance of Animals* (Univ. Chicago Press, Chicago, 1954).
14. Menge, B. A. & Sutherland, J. P. Species diversity gradients: synthesis of the role of predation, competition and temporal heterogeneity. *Am. Nat.* **110**, 351–369 (1976).
15. Schoener, T. W. in *Community and Evolutionary Ecology of North American Stream Fishes* (eds Matthews, W. J. & Heins, D. C.) 8–16 (Univ. Oklahoma Press, Norman, 1987).
16. Wiens, J. A. On competition and variable environments. *Am. Sci.* **65**, 590–597 (1977).
17. Diamond, J. & Case, T. J. *Community Ecology* (Harper & Row, New York, 1986).
18. Evans, E. W. Influence of weather on predator/prey relations: stinkbugs and tent caterpillars. *J. NY Entomol. Soc.* **90**, 241–246 (1982).
19. Holt, R. & Lawton, J. H. The ecological consequences of shared natural enemies. *Annu. Rev. Ecol. Syst.* **25**, 495–520 (1994).
20. Schoener, T. W. & Spiller, D. A. Devastation of prey diversity by experimentally introduced predators in the field. *Nature* **381**, 691–694 (1996).
21. Spiller, D. A., Losos, J. B. & Schoener, T. W. Impact of a catastrophic hurricane on island populations. *Science* **281**, 695–697 (1998).
22. Ackerman, J. New eyes on the ocean. *Nat. Geogr. Mag.* Oct. 86–114 (2000).
23. Kareiva, P. M., Kingsolver, J. G. & Huey, R. B. (eds) *Biotic Interactions and Global Change* (Sinauer, Sunderland, Massachusetts, 1993).
24. Schwartz, A. & Henderson, R. W. *Amphibians and Reptiles of the West Indies* 424 (Univ. Florida Press, Gainesville, 1991).
25. Schoener, T. W. Presence and absence of habitat shift in some widespread lizard species. *Ecol. Monogr.* **45**, 233–258 (1975).
26. Schoener, T. W. The *Anolis* lizards of Bimini: resource partitioning in a complex fauna. *Ecology* **49**, 704–726 (1968).
27. Heckel, D. G. & Roughgarden, J. A technique for estimating the size of lizard populations. *Ecology* **60**, 966–975 (1979).
28. Fienberg, S. E. The multiple-recapture census. *Biometrika* **45**, 591–603 (1972).
29. Siegel, S. & Castellan, N. J. Jr *Nonparametric Statistics for the Behavioral Sciences* 128–137 (McGraw-Hill, New York, 1988).

Acknowledgements

We thank the NSF for support.

Correspondence and requests for reprints should be addressed to T.W.S. (e-mail: twschoener@ucdavis.edu).

Essential role for Gab2 in the allergic response

Haihua Gu*, Kan Saito†, Lori D. Klamon*, Junqing Shen*, Tony Fleming†, YongPing Wang*, Joanne C. Pratt‡, Guosheng Lin*, Bing Lim*, Jean-Pierre Kinet† & Benjamin G. Neel*

* Cancer Biology Program, Division of Hematology and Oncology; and † Division of Experimental Pathology, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA

‡ Franklin W. Olin College of Engineering, 1753 Great Plain Avenue, Needham, Massachusetts 02492, USA

Dos/Gab family scaffolding adapters (Dos, Gab1, Gab2) bind several signal relay molecules, including the protein-tyrosine phosphatase Shp-2 and phosphatidylinositol-3-OH kinase (PI(3)K); they are also implicated in growth factor, cytokine and antigen receptor signal transduction¹. Mice lacking Gab1 die during embryogenesis and show defective responses to several stimuli^{2,3}. Here we report that Gab2^{-/-} mice are viable and generally healthy; however, the response (for example, degranulation and cytokine gene expression) of Gab2^{-/-} mast cells to stimulation of the high affinity immunoglobulin-ε (IgE) receptor FcεRI is defective. Accordingly, allergic reactions such as passive

cutaneous and systemic anaphylaxis are markedly impaired in *Gab2*^{-/-} mice. Biochemical analyses reveal that signalling pathways dependent on PI(3)K, a critical component of FcεRI signalling, are defective in *Gab2*^{-/-} mast cells. Our data identify Gab2 as the principal activator of PI(3)K in response to FcεRI activation, thereby providing genetic evidence that Dos/Gab family scaffolds regulate the PI(3)K pathway *in vivo*. Gab2 and/or its associated signalling molecules may be new targets for developing drugs to treat allergy.

FcεRI mediates the allergic response by inducing the release of vasoactive agents, including histamine and serotonin, from preformed granules (degranulation), as well as the synthesis and release of cytokines. Similar to lymphocyte antigen receptors and Fcγ receptors on phagocytes, FcεRI is a multi-chain immune recognition receptor (MIRR), which signals by activating receptor-associated Src and Syk family protein-tyrosine kinases (PTKs). On crosslinking of FcεRIα subunits, the Src family PTK Lyn phosphorylates the associated FcεRIβ and -γ signalling chains. This leads to the recruitment of Syk, which phosphorylates several signal relay molecules including LAT (linker for activation of T cells)⁴, and activates downstream signalling cascades⁵. An important intermediate is PI(3)K, which catalyses phosphatidylinositol-3,4,5-bisphosphate and phosphatidylinositol-3,4,5-trisphosphate (PtdIns(3,4,5)P₃) synthesis. These phospholipids recruit molecules containing Pleckstrin homology (PH) domains, including Tec PTKs⁶, the serine kinase Akt⁷ and phospholipase Cγ1 (PLCγ1)^{8,9}. PI(3)K inhibitors abrogate the mast cell calcium response required for degranulation^{10–12} and impair the activation of mitogen-activated protein kinases (MAPKs) needed for cytokine gene induction^{13,14}. However, the mechanism of PI(3)K activation in FcεRI signalling has remained unclear. Here we show that Gab2 is the missing link between FcεRI engagement and PI3K activation, and is essential for FcεRI-evoked degranulation, cytokine synthesis and anaphylaxis.

Gab2, like its relatives *Drosophila* DOS and mammalian Gab1, contains an amino-terminal PH domain and several binding sites for signal relay molecules, including Grb2, Shc, Shp-2 and the p85 subunit of PI(3)K. DOS and Gab1 are required for several pathways in embryogenesis^{2,3,15}. Like Gab1, Gab2 is expressed widely¹⁶. Although studies using cultured cells implicate it in growth factor, cytokine and antigen receptor signalling^{16–19}, the physiological function of Gab2 is unclear.

To address this issue we generated *Gab2*^{-/-} mice that were born at the expected mendelian frequency (Supplementary Information Fig. 1a, b, c) and remained viable and apparently healthy for up to 15 months. Notably, Gab2 protein was undetectable in several *Gab2*^{-/-} tissues assayed (Supplementary Information Fig. 1d). Histology of several organs (carried out on several mice within 6 months of age), B- and T-cell development (as assessed by flow cytometry), bone marrow cellularity, and peripheral blood counts were grossly normal in *Gab2*^{-/-} mice. Thus, Gab2 seems dispensable for lymphocyte development and steady-state haematopoiesis (data not shown).

Previous studies suggested that Gab2 is important for PI(3)K activation in response to cytokines that use the interleukin 3 (IL-3) receptor beta chain¹⁸. To test this in primary cells, we cultured bone-marrow-derived mast cells (BMMCs) from wild type and *Gab2*^{-/-} mice. IL-3-induced Akt activation was substantially diminished in the *Gab2*^{-/-} BMMCs (data not shown).

As biochemical data implicates Gab2 in MIRR signalling^{16,17,19}, we investigated whether Gab2 has a function in FcεRI responses. Wild-type BMMCs were sensitized with anti-dinitrophenol (DNP) IgE and stimulated with DNP. Gab2 became tyrosyl phosphorylated and associated with p85 and Shp-2 (Fig. 1a). Therefore, we compared degranulation in wild-type and *Gab2*^{-/-} BMMCs. FcεRI-evoked serotonin release was markedly inhibited in *Gab2*^{-/-} BMMCs (Fig. 1b). Similar defects were observed in *Gab2*^{-/-} mice derived from two embryonic stem cell clones and in mice on hybrid and

Sv129 backgrounds (data not shown). 12-*O*-tetradecanoylphorbol-13-acetate (TPA) plus ionomycin-induced degranulation was comparable in wild-type and *Gab2*^{-/-} BMMCs (Fig. 1c). Therefore, the secretion machinery is intact in *Gab2*^{-/-} BMMCs and defective signalling must be responsible for decreased FcεRI-evoked degranulation. FcεRI-evoked tumour-necrosis factor-α (TNF-α) and IL-6 messenger RNA synthesis were decreased by 50–65% in *Gab2*^{-/-} BMMCs (Fig. 1d, e). These defects cause a diminished allergic response in the whole animal: passive cutaneous (Fig. 1f) and passive systemic (Fig. 1g) anaphylaxis were reduced significantly in *Gab2*^{-/-} mice.

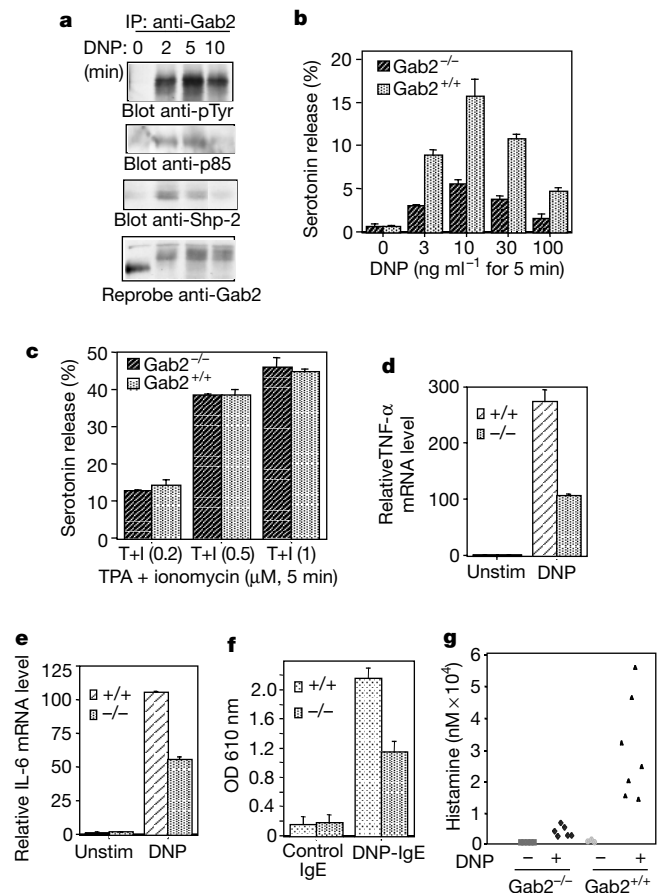


Figure 1 Gab2 is essential for FcεRI-evoked biological responses in BMMCs. **a**, Gab2 is tyrosyl phosphorylated and associates with p85 and Shp-2 on FcεRI activation. Wild-type (+/+) BMMCs were sensitized with anti-DNP-IgE and stimulated with DNP (10 ng ml⁻¹). Anti-Gab2 immunoprecipitates were immunoblotted with the indicated antibodies. **b, c**, FcεRI-mediated degranulation is impaired in *Gab2*^{-/-} BMMCs. BMMCs were sensitized as in **a**, labelled with ³H-serotonin and stimulated with DNP (**b**) or TPA plus ionomycin (**c**). Serotonin in the media and the cell pellet was quantified by scintillation counting. Data are expressed as the percentage of total counts (medium plus pellet) released and are the average of triplicate samples. A representative experiment (of four) is shown. **d, e**, FcεRI-induced cytokine gene expression is impaired in *Gab2*^{-/-} BMMCs. Sensitized BMMCs were stimulated with DNP (10 ng ml⁻¹) for 1 h or left unstimulated, and TNF-α (**d**) and IL-6 (**e**) transcripts were quantified by Real Time PCR. **f**, Decreased passive cutaneous anaphylaxis (PCA) in *Gab2*^{-/-} mice. A representative result (3 male mice per genotype) from two independent experiments. For unclear reasons, which may be related to hormonal effects on vascular tone, we observed variably high backgrounds in PCA assays carried out on female mice. **g**, Decreased passive systemic anaphylaxis in *Gab2*^{-/-} mice. Wild-type and *Gab2*^{-/-} mice were sensitized with anti-DNP-IgE and either challenged (+) with DNP (*n* = 7 and 6 for wild-type and *Gab2*^{-/-} mice, respectively) or left untreated (-) (*n* = 3 and 5 for wild-type and *Gab2*^{-/-} mice, respectively). Histamine release from *Gab2*^{-/-} mice is significantly less than release from wild-type mice (*P* < 0.01 by unpaired two-tailed *t*-test).

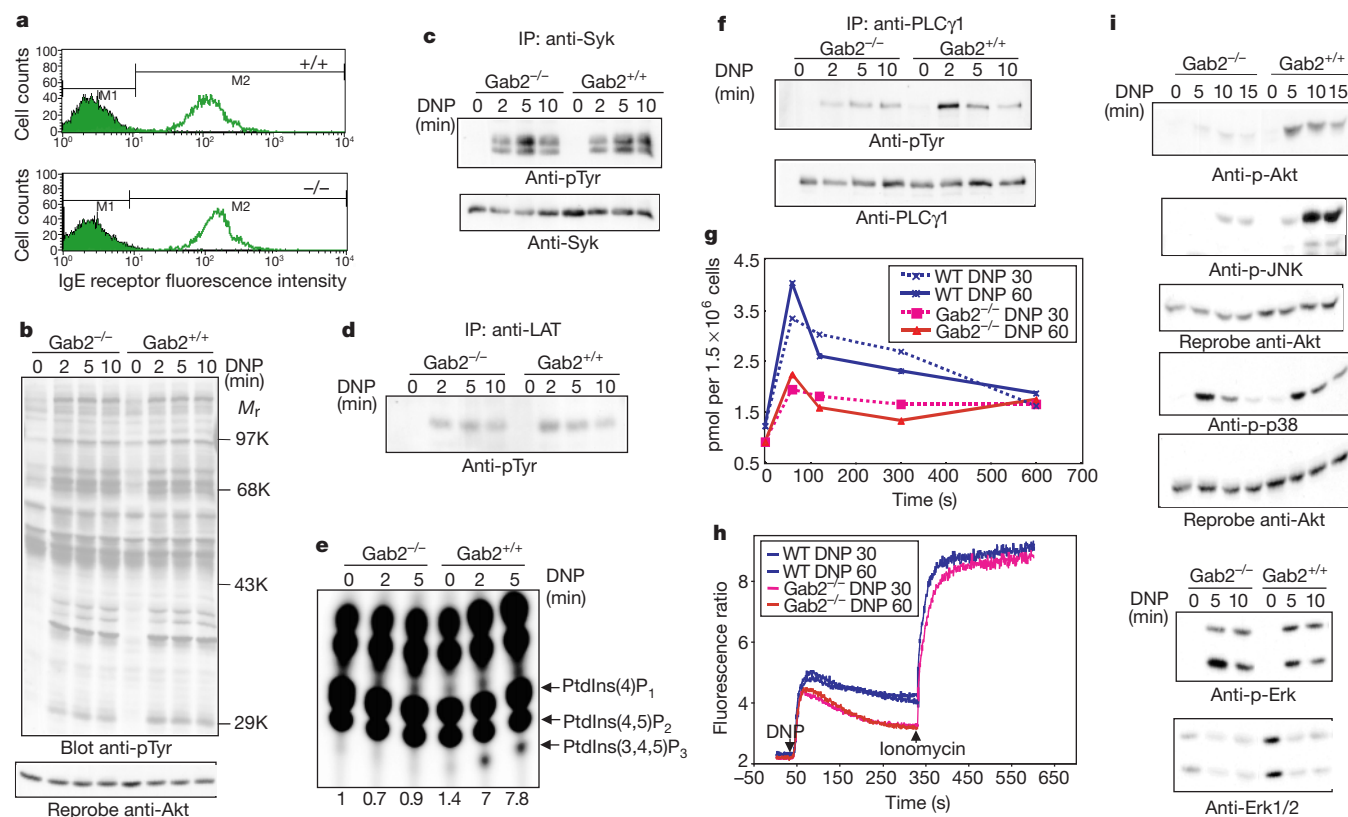


Figure 2 PI(3)K-dependent signalling is impaired in $Gab2^{-/-}$ BMMCs. **a**, Surface expression of FcεRI is normal in $Gab2^{-/-}$ BMMCs. **b–d**, FcεRI-induced total cellular (**b**), Syk (**c**) and LAT (**d**) tyrosyl phosphorylation are normal in $Gab2^{-/-}$ BMMCs. BMMCs were sensitized with anti-DNP-IgE and stimulated with 10 ng ml^{-1} DNP. Lysates were loaded directly (**b**) or immunoprecipitated with anti-Syk (**c**) or anti-LAT antibodies (**d**) and subjected to anti-phosphotyrosine immunoblotting (pTyr). In **b** the membrane was re-probed with anti-Akt antibodies; in **c** the blot was re-probed with anti-Syk antibodies. Experiments in **d** used equal, total protein and were repeated several times. **e**, FcεRI-induced PtdIns(3,4,5)P₃ production is decreased in $Gab2^{-/-}$ BMMCs. Lipids were extracted from sensitized BMMCs labelled with ^{32}P -orthophosphate and stimulated with DNP or left unstimulated and separated by thin-layer chromatography. The relative

intensity of each PtdIns(3,4,5)P₃ spot (quantified by densitometry) is shown. There was no significant FcεRI-stimulated PtdIns(3,4,5)P₃ production in $Gab2^{-/-}$ BMMCs. In other experiments small increases were observed. **f**, FcεRI-induced PLCγ1 phosphorylation is impaired in $Gab2^{-/-}$ BMMCs. **g**, FcεRI-evoked IP3 production is impaired in $Gab2^{-/-}$ BMMCs. Sensitized BMMCs were stimulated with DNP (30 or 60 ng ml^{-1}) and IP3 production was measured. **h**, FcεRI-evoked calcium mobilization is defective in $Gab2^{-/-}$ BMMCs. Sensitized BMMCs were loaded with Fura-2/AM and stimulated with DNP (30 and 60 ng ml^{-1}) or ionomycin (2 μM). Calcium flux was monitored by spectrofluorimetry. **i**, FcεRI-induced Akt, JNK and p38 activation are impaired in $Gab2^{-/-}$ BMMCs. Lysates were immunoblotted with phospho-specific antibodies for Akt (S473), JNK, p38 and Erk, followed by re-probing with anti-Akt antibodies to control for loading.

We next monitored FcεRI-evoked signalling events. FcεRI expression (Fig. 2a) and total cellular, Syk and LAT tyrosyl phosphorylation (Fig. 2b, c, d) were normal in $Gab2^{-/-}$ BMMCs. Thus, Gab2 must function downstream of, or parallel to, LAT phosphorylation. As PI(3)K binds to Gab2 (Fig. 1a; see also ref. 18), we asked whether Gab2 is important for PI(3)K lipid product generation. FcεRI-evoked PtdIns(3,4,5)P₃ levels were reduced by over 80% in $Gab2^{-/-}$ BMMCs (Fig. 2e), establishing Gab2 as a critical regulator of PI(3)K in the FcεRI pathway. PLCγ1 tyrosyl phosphorylation was also reduced in $Gab2^{-/-}$ BMMCs (Fig. 2f). PI(3)K lipid products are important for optimal recruitment of PLCγ to cell membranes^{8,9,20}, as well as for the recruitment and activation of Tec family PTKs, implicated in PLCγ phosphorylation⁶. Tyrosyl-phosphorylated PLCγ1 catalyses the generation of diacylglycerol (DAG) and inositol 1,4,5-trisphosphate (IP3), and IP3 is required for calcium release from internal stores. Accordingly, FcεRI-evoked IP3 production (Fig. 2g) and calcium release were impaired in $Gab2^{-/-}$ BMMCs (Fig. 2h). PI(3)K is implicated in the activation of several serine kinase cascades downstream from FcεRI. Consistent with their decreased PtdIns(3,4,5)P₃ levels, Akt activation was markedly diminished in $Gab2^{-/-}$ BMMCs (Fig. 2i). JNK and to a lesser extent p38 activation were also reduced in $Gab2^{-/-}$ BMMCs, whereas Erk activation was largely unaffected.

Sections stained with Toluidine blue revealed similar numbers of

mast cells in the ear skins of wild-type and $Gab2^{-/-}$ mice, but mast cell numbers in back and thigh skins were reduced by 30–50%, whereas peritoneal mast cells were undetectable in $Gab2^{-/-}$ mice (data not shown). Gab2 may therefore have an additional function in mast cell ontogeny, probably in the stem-cell factor (SCF)/c-Kit pathway²¹. To exclude the possibility that the functional defects were indirect consequences of defective mast cell development, we stably reconstituted Gab2 expression in $Gab2^{-/-}$ BMMCs by using a retroviral vector. Re-expression of Gab2 rescued the defect in FcεRI-evoked degranulation in $Gab2^{-/-}$ BMMCs (Fig. 2a). Compared with cells infected with a control virus, Gab2-reconstituted BMMCs showed markedly increased Akt, JNK and p38 activation in response to FcεRI engagement (Fig. 3b). Importantly, the ability of Gab2 to rescue FcεRI-evoked Akt activation required intact Gab2 binding sites for the p85 subunit of PI(3)K, indicating that FcεRI-evoked PI(3)K activation requires direct recruitment of PI(3)K by Gab2 (Fig. 3c).

FcεRI-evoked responses and signalling events are not eliminated totally in $Gab2^{-/-}$ BMMCs. Although we detected no Gab2 protein in several tissues (Supplementary Information Fig. 1d; and data not shown), $Gab2^{-/-}$ BMMCs expressed small (~15% of wild type) amounts of a truncated form of Gab2 (Fig. 3d). Preliminary data suggest that this protein is encoded by a transcript arising from a cryptic promoter in the first intron of the Gab2 gene. The truncated

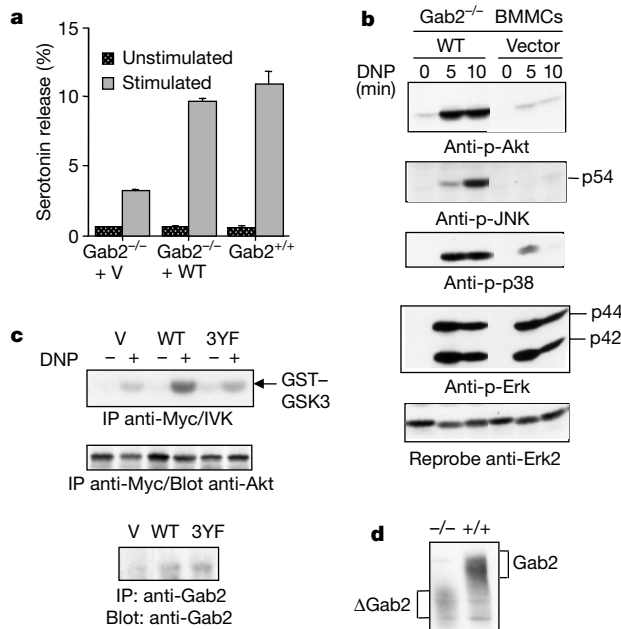


Figure 3 Restoring Gab2 expression rescues FcεRI-evoked degranulation and signalling defects in Gab2^{-/-} BMMCs. Gab2^{-/-} BMMCs were infected with a retroviral vector co-expressing Gab2 and GFP (wild type, WT) or with Vector alone, and infected cells isolated by FACS. **a**, Re-expression of Gab2 wild type in Gab2^{-/-} BMMCs rescues defective degranulation. **b**, Re-expression of Gab2 wild type in Gab2^{-/-} BMMCs rescues defective signalling. Cells were stimulated through FcεRI or left unstimulated; activation of the indicated kinases was assessed as in Fig. 2i. **c**, p85 binding to Gab2 is required for FcεRI-evoked Akt activation. Gab2^{-/-} BMMCs were transiently co-transfected with Myc-Akt and Gab2 wild type or Gab2 3YF expression vectors, or empty vector. Transfected cells were stimulated as in Fig. 2i. Myc-Akt immunoprecipitates were subjected to *in vitro* kinase assays using GST-GSK3. Reprobing with anti-Akt antibodies shows comparable recovery Myc-Akt. Expression of the transfected (haemagglutinin (HA)-tagged) Gab2 wild type and Gab2 3YF in Gab2^{-/-} BMMCs is shown by Gab2 immunoprecipitation/immunoblotting. **d**, Gab2^{-/-} BMMCs express small amounts of truncated Gab2. Lysates were immunoblotted with anti-Gab2 antibodies. Gab2^{-/-} BMMCs express a truncated (Δ) Gab2 protein at about 15% the level of full-length Gab2 protein in wild-type BMMCs.

protein should lack the first 39 amino acids of the Gab2 PH domain, including the β1–β2 loop, important for binding phosphatidylinositol lipids²². Experiments similar to those in Fig. 3c show that the truncated protein retains some ability to promote FcεRI-evoked Akt activation, but only when expressed at levels roughly 50-fold above that of the truncated endogenous protein produced in Gab2^{-/-} BMMCs (data not shown). Although we cannot exclude the possibility that truncated Gab2 contributes to residual FcεRI function in Gab2^{-/-} BMMCs, this is unlikely. The low expression of the truncation mutant also argues against significant dominant negative effects or gratuitous actions on other signalling pathways. More likely, other, less effective pathways to PI(3)K activation may exist. Such a pathway(s) could emanate from Cbl²³, CD19 (ref. 24) or an as-yet-undiscovered Gab family member (or other scaffolding adapter) expressed at lower levels in BMMCs. Notably, we cannot detect Gab1 in BMMCs (data not shown).

Our results show conclusively that Gab2 is an essential, cell autonomous signalling molecule in FcεRI signal transduction (Fig. 4). A central function of Gab2 is to activate PI(3)K by binding its p85 subunit and recruiting it to its substrate lipids. Decreased JNK and p38 activity might reflect impaired recruitment of guanine nucleotide exchange proteins containing PH domains (for example Vav) for small G proteins (such as Rac2); Rac2^{-/-} (ref. 25) BMMCs also have impaired FcεRI responses. Defective cytokine gene

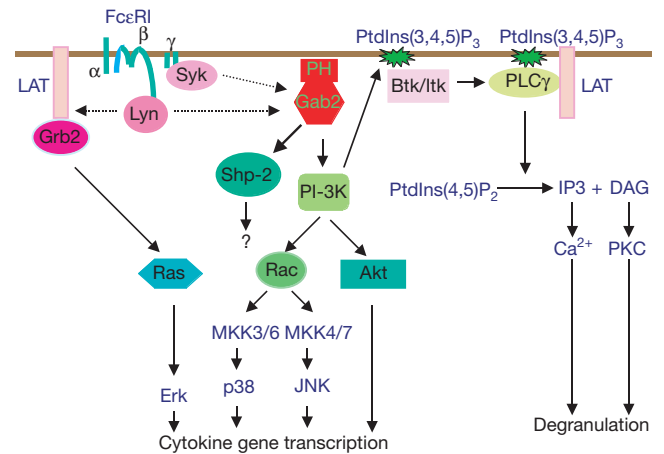


Figure 4 Model of Gab2 action in FcεRI signalling. On FcεRI stimulation, Gab2 becomes tyrosyl phosphorylated, probably by Syk or Lyn. Phosphorylated Gab2 binds PI(3)K through p85, as well as other signal relay molecules, such as Shp-2. Gab-2/PI(3)K complexes account for most PtdIns(3,4,5)P₃ generation on FcεRI stimulation. PtdIns(3,4,5)P₃ is required for Tec family kinase activation, and subsequent phosphorylation and activation of PLCγ1, as well as for activation of Akt, JNK and to a lesser extent, p38. The role of other signalling molecules that bind Gab2 remains to be defined.

induction with normal Erk activation in Gab2^{-/-} BMMCs (Fig. 1c, d) is reminiscent of our previous studies of Gab2 in cytokine signalling, in which Gab2 by means of Shp-2 functions in a pathway required for immediate-early gene induction, but not Erk activation¹⁶ (Fig. 2c, d). Decreased JNK and p38 activation may also contribute to lower cytokine gene transcription. Reconstitution studies using other phosphorylation site mutants of Gab2 should resolve such questions. Gab1^{-/-} fibroblasts demonstrate defective Erk activation in response to several growth factors and cytokines. These data, along with the markedly different phenotypes of Gab1^{-/-} (refs 2, 3) and Gab2^{-/-} mice, indicate that the two mammalian Gab proteins have distinct functions.

Identification of Gab2 as the central mediator of PI(3)K activation from FcεRI has more general implications. For example, the mechanism by which other MIRRs (such as FcγR and T-cell antigen receptor) activate PI(3)K has not been established clearly. Our data indicate that Gab family proteins are possible intermediates. The highly selective inhibition of immediate and delayed responses to IgE receptor triggering in Gab2^{-/-} BMMCs raises the possibility that strategies to lower Gab2 levels might have therapeutic benefits in atopic individuals. □

Methods

Generation of Gab2^{-/-} mice

We generated Gab2^{-/-} mice by targeting exon 1 using a standard replacement strategy. (See Supplementary Information for details of generation and characterization of these mice.)

Antibodies and reagents

Monoclonal fluorescein isothiocyanate (FITC)-anti-mouse IgE was provided by H. Katz. Monoclonal anti-DNP mouse IgE, SPE-7 and DNP-HSA were from Sigma. Polyclonal anti-Gab2 and anti-p85 antibodies were described previously¹⁶. Anti-PLCγ1, anti-Akt and anti-Shp-2 were purchased from Santa Cruz. Monoclonal anti-phosphotyrosine antibody 4G10 was from UBI, and anti-phospho-p44/p42, anti-phospho-JNK, anti-phospho-p38 and anti-phospho-Akt (Ser 473) were from Cell Signaling. Anti-Syk antibodies were provided by J. Cambier.

Mast cell cultures and functional assays

Murine bone marrow was cultured in IMDM containing 10% FCS and 2–4 ng ml⁻¹ recombinant murine IL-3 (Intergen). We used BMMCs cultured for 4–7 weeks for all studies. FcεRI surface expression was quantified by flow cytometry as described²⁶. We measured degranulation by serotonin release as described⁴. To determine cytokine mRNA levels, BMMCs were sensitized with IgE (2 μg ml⁻¹ for 4 h), washed, resuspended in Tyrode's buffer, and left unstimulated or stimulated with DNP (10 ng ml⁻¹) for 1 h. Total

RNA was isolated using TRIzol (GIBCO). First-strand complementary DNA synthesized from total RNA using Superscript reverse transcriptase (GIBCO) was subjected to Real Time PCR using a Taqman instrument (Perkin–Elmer). The relative levels of cytokine mRNAs in each sample were normalized to GAPDH levels. We performed passive cutaneous anaphylaxis²⁷ and passive systemic anaphylaxis (PCA and PSA, respectively) assays essentially as described⁴. We used 2.5–3-month-old mice for these experiments.

Biochemical analyses

Lysis, immunoprecipitation and immunoblotting were carried out as described¹⁶, except that 1% Triton was used in place of NP-40. PtdIns(3,4,5)P₃ levels were measured by thin-layer chromatography as described⁶. IP3 concentration was assayed using a kit from NEN.

BMMC calcium responses

BMMCs were sensitized with anti-IgE (2 µg ml⁻¹) for 4 h, loaded with 2 µM Fura-2/AM (Molecular Probes) in buffer also containing 0.25 mM sulphapyrazone at 37 °C for 25 min, washed, and stimulated with DNP at the indicated concentrations. Intracellular calcium measurements were performed using a Deltascan spectrofluorimeter as described⁶.

BMMC reconstitution assays

Full-length Gab2 cDNA was cloned into the EcoRI site of the retroviral vector MSCV–IRES–green fluorescent protein (GFP), generating MSCV–IRES–GFP–Gab2. Viral supernatants were produced as described²⁸, and used to infect three-week-old BMMC cultures containing polybrene (4 µg ml⁻¹) by spin infection (800 g, room temperature for 90 min). Seven days after infection, GFP-positive cells were purified by FACS and cultured for 10 d before analysis. For transient reconstitutions, BMMCs (1.5 × 10⁷) were electroporated with 3 µg Myc–Akt plasmid and 18 µg pEBB, pEBB Gab2 wild type or pEBB Gab2 3YF at 380 V and 960 µFD using a Bio-Rad GenePulser. Twenty-four hours after electroporation, cells were sensitized and stimulated with DNP. Myc–Akt immunoprecipitates were subjected to *in vitro* kinase assays using glutathione S-transferase (GST)–GSK3 as the substrate²⁹.

Received 20 March; accepted 15 May 2001.

- Hibi, M. & Hirano, T. Gab-family adapter molecules in signal transduction of cytokine and growth factor receptors, and T and B cell antigen receptors. *Leuk. Lymphoma* **37**, 299–307 (2000).
- Itoh, M. *et al.* Role of Gab1 in heart, placenta, and skin development and growth factor- and cytokine-induced extracellular signal-regulated kinase mitogen-activated protein kinase activation. *Mol. Cell Biol.* **20**, 3695–3704 (2000).
- Sachs, M. *et al.* Essential role of Gab1 for signaling by the c-Met receptor *in vivo*. *J. Cell Biol.* **150**, 1375–1384 (2000).
- Saitoh, S. *et al.* LAT is essential for Fc(ε)RI-mediated mast cell activation. *Immunity* **12**, 525–535 (2000).
- Turner, H. & Kinet, J. P. Signalling through the high-affinity IgE receptor FcεRI. *Nature* **402**, B24–30 (1999).
- Scharenberg, A. M. *et al.* Phosphatidylinositol-3,4,5-trisphosphate (PtdIns-3,4,5-P₃)/Tec kinase-dependent calcium signaling pathway: a target for SHIP-mediated inhibitory signals. *EMBO J.* **17**, 1961–1972 (1998).
- Franke, T. F., Kaplan, D. R., Cantley, L. C. & Tokar, A. Direct regulation of the Akt proto-oncogene product by phosphatidylinositol-3,4-bisphosphate. *Science* **275**, 665–668 (1997).
- Falasca, M. *et al.* Activation of phospholipase Cγ by PI 3-kinase-induced PH domain-mediated membrane targeting. *EMBO J.* **17**, 414–422 (1998).
- Bae, Y. S. *et al.* Activation of phospholipase C-γ by phosphatidylinositol 3,4,5-trisphosphate. *J. Biol. Chem.* **273**, 4465–4469 (1998).
- Nakanishi, S., Yano, H. & Matsuda, Y. Novel functions of phosphatidylinositol 3-kinase in terminally differentiated cells. *Cell Signal* **7**, 545–557 (1995).
- Barker, S. A. *et al.* Wortmannin blocks lipid and protein kinase activities associated with PI 3-kinase and inhibits a subset of responses induced by FcεRI cross-linking. *Mol. Biol. Cell* **6**, 1145–1158 (1995).
- Barker, S. A., Lujan, D. & Wilson, B. S. Multiple roles for PI 3-kinase in the regulation of PLCγ activity and Ca²⁺ mobilization in antigen-stimulated mast cells. *J. Leukoc. Biol.* **65**, 321–329 (1999).
- Kawakami, Y., Hartman, S. E., Holland, P. M., Cooper, J. A. & Kawakami, T. Multiple signalling pathways for the activation of JNK in mast cells: involvement of Bruton's tyrosine kinase, protein kinase C, and JNK kinases, SEK1 and MKK7. *J. Immunol.* **161**, 1795–1802 (1998).
- Ishizuka, T. *et al.* Mitogen-activated protein kinase activation through Fcε receptor I and stem cell factor receptor is differentially regulated by phosphatidylinositol 3-kinase and calcineurin in mouse bone marrow-derived mast cells. *J. Immunol.* **162**, 2087–2094 (1999).
- Raabe, T. *et al.* DOS, a novel pleckstrin homology domain-containing protein required for signal transduction between sevenless and Ras1 in *Drosophila*. *Cell* **85**, 911–920 (1996).
- Gu, H., Pratt, J. C., Burakoff, S. J. & Neel, B. G. Cloning of p97/Gab2, the major SHP-2 binding protein in hematopoietic cells, reveals a novel pathway for cytokine-induced gene activation. *Mol. Cell* **2**, 729–740 (1998).
- Nishida, K. *et al.* Gab-family adapter proteins act downstream of cytokine and growth factor receptors and T- and B-cell antigen receptors. *Blood* **93**, 1809–1816 (1999).
- Gu, H. *et al.* New role for Shc in activation of the phosphatidylinositol 3-kinase/Akt pathway. *Mol. Cell Biol.* **20**, 7109–7120 (2000).
- Pratt, J. C. *et al.* Cutting edge: gab2 mediates an inhibitory phosphatidylinositol 3'-kinase pathway in T cell antigen receptor signaling. *J. Immunol.* **165**, 4158–4163 (2000).
- Rameh, L. E. *et al.* Phosphoinositide 3-kinase regulates phospholipase Cγ-mediated calcium signaling. *J. Biol. Chem.* **273**, 23750–23757 (1998).
- Galli, S. J. Mast cells and basophils. *Curr. Opin. Hematol.* **7**, 32–39 (2000).
- Rameh, L. E. *et al.* A comparative analysis of the phosphoinositide binding specificity of pleckstrin homology domains. *J. Biol. Chem.* **272**, 22059–22066 (1997).

- Liu, Y. C. & Altman, A. Cbl: complex formation and functional implications. *Cell Signal* **10**, 377–385 (1998).
- Gommerman, J. L. *et al.* A role for CD21/CD35 and CD19 in responses to acute septic peritonitis: a potential mechanism for mast cell activation. *J. Immunol.* **165**, 6915–6921 (2000).
- Yang, F. C. *et al.* Rac2 stimulates Akt activation affecting BAD/Bcl-XL expression while mediating survival and actin function in primary mast cells. *Immunity* **12**, 557–568 (2000).
- Lu-Kuo, J. M., Fruman, D. A., Joyal, D. M., Cantley, L. C. & Katz, H. R. Impaired kit- but not FcεRI-initiated mast cell activation in the absence of phosphoinositide 3-kinase p85α gene products. *J. Biol. Chem.* **275**, 6022–6029 (2000).
- Dombrowicz, D. *et al.* Absence of FcεRI alpha chain results in upregulation of FcγRIII-dependent mast cell degranulation and anaphylaxis. Evidence of competition between FcεRI and FcγRIII for limiting amounts of Fcβ and gamma chains. *J. Clin. Invest.* **99**, 915–925 (1997).
- Schwaller, J. *et al.* Transformation of hematopoietic cell lines to growth-factor independence and induction of a fatal myelo- and lymphoproliferative disease in mice by retrovirally transduced TEL/JAK2 fusion genes. *EMBO J.* **17**, 5321–5333 (1998).
- Craddock, B. L., Orchiston, E. A., Hinton, H. J. & Welham, M. J. Dissociation of apoptosis from proliferation, protein kinase B activation, and BAD phosphorylation in interleukin-3-mediated phosphoinositide 3-kinase signaling. *J. Biol. Chem.* **274**, 10633–10640 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. Lawitts for blastocyst injections; V. Petkova for help with Real Time PCR analysis; D. P. Liu for technical assistance; and L. C. Cantley and C. Carpenter for comments. This work was supported by grants from the N.I.H. (B.G.N., J.-P.K. and B.L.). H.G. was supported by fellowships from the N.I.H. and the American Association for Cancer Research. J.C.P. was supported by grants from the Cancer League of Colorado and the Association for International Cancer Research.

Correspondence and requests for materials should be addressed to H.G. (e-mail: hgu@caregroup.harvard.edu).

The heparin-binding haemagglutinin of *M. tuberculosis* is required for extrapulmonary dissemination

Kevin Pethe*, Sylvie Alonso*, Franck Biet*, Giovanni Delogu†, Michael J. Brennan†, Camille Locht* & Franco D. Menozzi*

* Unité INSERM U447, Institut Pasteur de Lille, 1 rue du Prof. Calmette, F-59019 Lille Cedex, France

† Laboratory of Mycobacterial Diseases, Center for Biologics Evaluation and Research, Food and Drug Administration, 29 Lincoln Drive, Building 29, Room 502, Bethesda, Maryland 20892, USA

Tuberculosis remains the world's leading cause of death due to a single infectious agent, *Mycobacterium tuberculosis*, with 3 million deaths and 10 million new cases per year¹. The infection initiates in the lungs and can then spread rapidly to other tissues². The availability of the entire *M. tuberculosis* genome sequence³ and advances in gene disruption technologies⁴ have led to the identification of several mycobacterial determinants involved in virulence^{5–8}. However, no virulence factor specifically involved in the extrapulmonary dissemination of *M. tuberculosis* has been identified to date. Here we show that the disruption of the *M. tuberculosis* or *Mycobacterium bovis* Bacille Calmette–Guérin (BCG) *hbhA* gene encoding the heparin-binding haemagglutinin adhesin (HBHA) markedly affects mycobacterial interactions with epithelial cells, but not with macrophage-like cells. When nasally administered to mice, the mutant strains were severely impaired in spleen colonization, but not in lung colonization. Coating wild-type mycobacteria with anti-HBHA antibodies also impaired dissemination after intranasal infection. These results provide evidence that adhesins such as HBHA are required for extrapulmonary dissemination, and that interactions with non-phagocytic cells have an important role in the pathogenesis of

tuberculosis. They also suggest that antibody responses to HBHA may add to immune protection against tuberculosis.

Alveolar macrophages have long been considered to be the port of entry of *M. tuberculosis*⁹, and are thought to convey the engulfed bacteria from the lungs to other organs. However, *M. tuberculosis* has also been shown to interact with epithelial cells, including the pulmonary M cells¹⁰, which may allow the bacilli to directly cross the epithelial cell layer. The relative contribution of either mechanism as well as the bacterial factors involved in *M. tuberculosis* dissemination are unknown. *Mycobacterium tuberculosis* complex strains produce at their surface a heparin-binding haemagglutinin adhesin (HBHA)¹¹. This protein is also produced by other pathogenic mycobacteria such as *Mycobacterium leprae* and *Mycobacterium avium*¹² but not by non-pathogenic *Mycobacterium smegmatis*¹³, a fast-growing saprophyte. Binding of *M. tuberculosis* to epithelial cells, but not to macrophages, can be inhibited with anti-HBHA antibodies or by competition with heparin¹¹, suggesting that HBHA-mediated adherence is specific for non-phagocytic cells. This adherence relies on the recognition of heparan sulphate-containing receptors by the carboxy-terminal lysine-rich domain of HBHA^{14,15}.

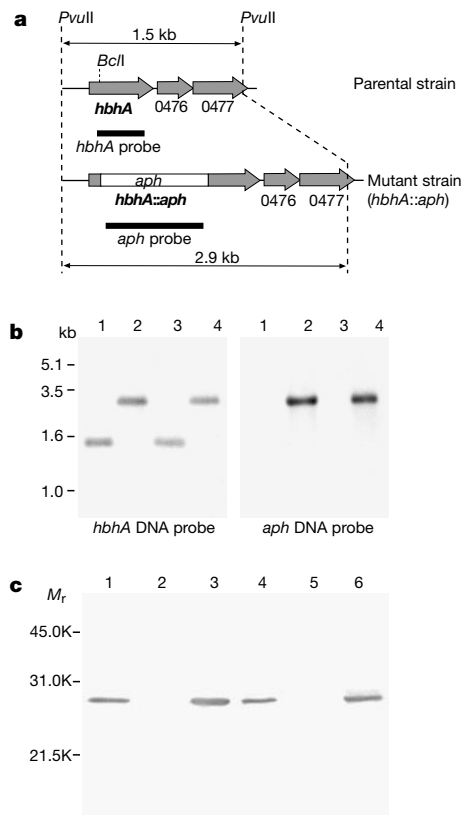


Figure 1 Genotype and phenotype analyses of *hbhA::aph* strains. **a**, Chromosomal organization of the *hbhA* gene in the parental and mutant strains of *M. bovis* BCG and *M. tuberculosis*. The arrows depict the lengths and directions of the *hbhA* gene and of open reading frames Rv0476 and Rv0477, as indicated. Black bars correspond to the probes used for Southern blot analysis. **b**, Southern blot analysis of chromosomal DNA from parental and mutant strains of *M. bovis* BCG (lanes 1 and 2) and *M. tuberculosis* (lanes 3 and 4). The probes used for hybridization are shown in **a**. **c**, Immunodetection of HBHA in cell-associated fractions of *M. bovis* BCG (lanes 1–3) and *M. tuberculosis* 103 (lanes 4–6) using anti-HBHA polyclonal antibodies described in ref. 13. Protein extracts were prepared from parental (lanes 1 and 4), mutant (lanes 2 and 5) and complemented (lanes 3 and 6) strains, and 50 µg total soluble proteins present in each extract was loaded onto the polyacrylamide gel.

To determine the role of HBHA in the mycobacterial interaction with non-phagocytic cells and in infection, a targeted null mutation was introduced at the *hbhA* locus¹⁶ of *M. tuberculosis* 103 and of the vaccine strain *M. bovis* BCG. Both mutants were constructed using *ts-SacB* technology⁴. The mutant strains resulting from allelic exchange are referred to here as MT103 *hbhA::aph* and BCG *hbhA::aph* (Fig. 1a). The disruption of *hbhA* was verified by Southern blot analysis using an *hbhA*- or an *aph*-specific DNA probe (Fig. 1b). This analysis revealed that both mutant strains contain a 1.4-kilobase (kb) insert at the *hbhA* locus, corresponding to the inserted *aph* gene. Both mutants were complemented by introducing a pYUB415 (ref. 17) derivative harbouring the wild-type *hbhA* gene under the control of its own promoter¹⁶. Western blot analysis of whole-cell lysates of the *hbhA::aph* strains using anti-HBHA antiserum demonstrated that the disruption of the *hbhA* gene resulted in the lack of HBHA production, which could be restored to wild-type levels by complementation (Fig. 1c).

The parental, *hbhA*-disrupted and complemented mutant strains showed similar colony morphology on Middlebrook 7H10 agar plates. When grown in Middlebrook 7H9/ADC broth or minimal Sauton medium, all strains exhibited identical doubling time and growth characteristics, indicating that HBHA is not required for the *in vitro* growth of *M. bovis* BCG or *M. tuberculosis* 103.

To determine whether *hbhA* disruption affects the mycobacterial interaction with eukaryotic cells, we evaluated the capacity of BCG *hbhA::aph* and MT103 *hbhA::aph* to adhere to, invade and survive within epithelial cells and macrophage-like cells. *In vitro* assays using J774.1, MH-S or U937 macrophage-like cell lines showed that HBHA is not involved in the *M. tuberculosis* interactions with these phagocytic cells, as adherence, invasion and intracellular multiplication of the *hbhA::aph* strain were similar to those of the parental and complemented strains (Fig. 2a, c). However, when human type II A549 pneumocytes were used as target cells, the adherence of the *hbhA::aph* strain was reduced by 60% compared with that of the

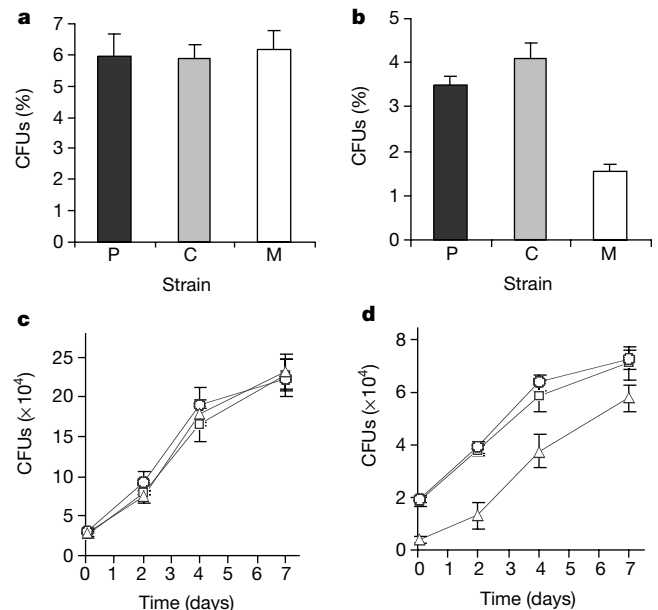


Figure 2 Effect of the *hbhA::aph* mutation on the interaction of *M. tuberculosis* 103 with phagocytic and non-phagocytic cells. **a**, **b**, Adherence of parental (P), complemented (C) and mutant (M) strains was assayed using the J774.1 macrophage-like cell line (**a**) and A549 pneumocytes (**b**), after 1 h. **c**, **d**, Invasion and survival of parental (squares), mutant (triangles) and complemented (circles) strains within J774.1 (**c**) cells and A549 pneumocytes (**d**). All assays were performed using an MOI of 10 mycobacteria per eukaryotic cell. Results are expressed as means of four different assays, and standard deviations are shown.

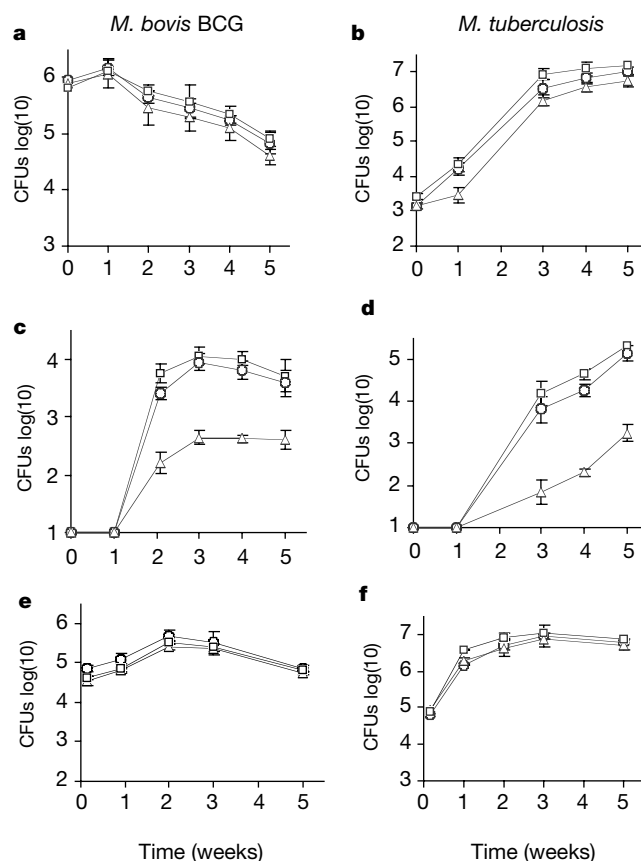


Figure 3 Effect of the *hbaA::aph* mutation on survival and multiplication of *M. bovis* BCG (left panels) and *M. tuberculosis* 103 (right panels). **a–f**, Effect of *hbaA::aph* mutation in lungs (**a** and **b**) and spleen (**c–f**) of BALB/c mice after intranasal infection with 10^3 and 10^6 viable units of *M. tuberculosis* 103 and *M. bovis* BCG, respectively (**a–d**), or

intravenous infection with 10^6 viable units of each strain (**e** and **f**). CFU counts are shown for the parental (squares), mutant (triangles) and complemented (circles) strains. The results for each point are the means and standard deviations of CFUs from four mice.

parental and complemented strains (Fig. 2b). Invasion of the pneumocytes by the *hbaA::aph* strain was reduced by 80%, as determined 1 h after infection (Fig. 2d). This reduction is probably a direct consequence of the reduced adhesiveness of the mutant strain, as its subsequent kinetics of intracellular growth were similar to those of the parental and complemented strains (Fig. 2d). Similar results were obtained for BCG strains (data not shown). Taken together, these observations indicate that HBHA is a principal mycobacterial adhesin for epithelial cells.

To determine whether HBHA has a function in mycobacterial infection, BALB/c mice were infected by the intranasal route with the parental, mutant or complemented strains of *M. tuberculosis* 103 or BCG. Infections were monitored by counting colony-forming units (CFUs) in the lungs and spleens of the infected mice over a period of 5 weeks. As shown in Fig. 3, the lung colonization profiles of BCG *hbaA::aph* (Fig. 3a) and MT103 *hbaA::aph* (Fig. 3b) were similar to those of the parental and complemented strains, although for MT103 *hbaA::aph*, a slight delay in lung colonization was observed during the first week after infection. These results indicate that HBHA does not have a large involvement in the initial infection or in mycobacterial persistence in the lungs. In contrast, the *hbaA::aph* strains were severely impaired in their ability to colonize the spleen. Three weeks after infection, the numbers of bacteria recovered from the spleens of the infected mice were 30-fold (BCG *hbaA::aph*, Fig. 3c) and 200-fold (MT103 *hbaA::aph*, Fig. 3d) lower compared with those of the parental strains. These results suggest that HBHA is crucial either for the mycobacteria to colonize and survive in the spleen or for extrapulmonary mycobacterial dissemination. To discriminate between these two possibilities, mice

were infected intravenously with the parental, mutant or complemented strains. As shown in Fig. 3e and f, the *hbaA::aph* strains colonized the spleen and persisted as well as the parental or complemented strains. This demonstrates that HBHA is not required for colonization and survival in the spleen, and that HBHA is therefore critical for escaping from the lungs and thus for extrapulmonary dissemination.

Further evidence for this hypothesis was obtained by coating mycobacteria with antibodies directed against HBHA¹¹, before intranasal administration of mice followed by monitoring lung and spleen colonizations. No significant difference in lung colonization was found between BCG coated with the anti-HBHA monoclonal antibodies 4057 D2 or 3921 E4, and BCG coated with a control monoclonal antibody or BCG alone (Fig. 4a). However, coating the BCG with the anti-HBHA antibodies led to a marked reduction in spleen colonization compared with uncoated BCG or BCG coated with the control antibody (Fig. 4b). As coating the mycobacteria with the anti-HBHA antibodies did not affect the numbers of the bacteria found in the lungs, there is no indication that the observed effect on dissemination is the result of a reduction in bacterial numbers due to differences in opsonization and resultant killing by alveolar macrophages. These observations suggest that mycobacterial extrapulmonary dissemination can be the result of a direct interaction with non-phagocytic cells without the need for an infection initially established in lung macrophages. However, it is still possible that after infection with low bacterial numbers, such as is observed in most natural *M. tuberculosis* infections in humans, an initial multiplication occurs in the lung macrophages before interaction with non-phagocytic cells and

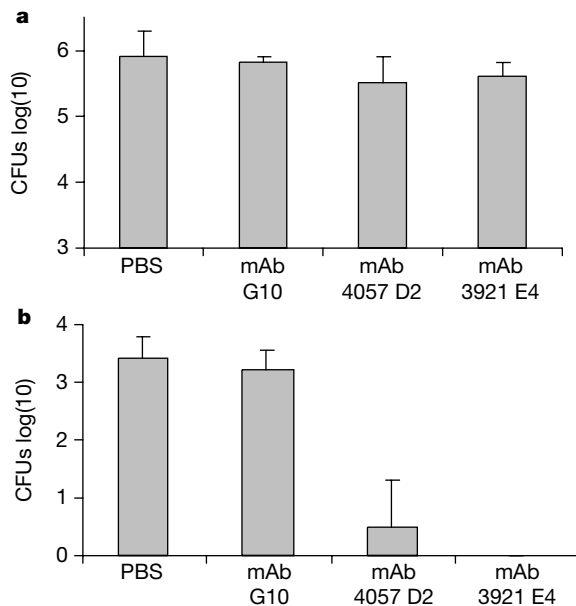


Figure 4 Influence of anti-HBHA antibodies on dissemination. Inocula of 5×10^6 *M. bovis* BCG pre-incubated with PBS, control monoclonal antibody G10, or with anti-HBHA monoclonal antibodies 4057 D2 or 3921 E4 were administered intranasally to C57BL/6 mice. **a**, **b**, CFUs were enumerated at 7 d after homogenization of lung (**a**) and spleen (**b**) tissues. Results indicate the means and standard deviations of CFUs from three mice per group. mAb, monoclonal antibody.

subsequent dissemination. In addition to supporting the view that HBHA has an important role in promoting mycobacterial dissemination, these results suggest that certain antibody responses may add to the protection against *M. tuberculosis*, as previously proposed¹⁸.

Our study indicates that the *M. tuberculosis* HBHA is a new virulence factor involved in mycobacterial dissemination from the lungs to the spleen. As HBHA specifically binds to non-phagocytic cells such as epithelial cells it follows that one of the crucial steps in disseminated infection involves mycobacterial interactions with non-phagocytic cells, a mechanism that has been overlooked in favour of the prominent *M. tuberculosis*–macrophage interactions. Nevertheless, epithelial cells, including type II pneumocytes¹³ and M cells¹⁰, have previously been considered as possible ports of entry of *M. tuberculosis* in addition to macrophages. However, the relative contribution of non-phagocytic target cells to disseminated infection could not be previously evaluated. As shown in this study, binding to non-phagocytic cells seems to be a major path to dissemination. Although the precise identity of the non-phagocytic cells and the HBHA receptors involved in dissemination remains unknown, this work constitutes an insight into the molecular understanding of a major step in the pathogenesis of tuberculosis and perhaps other mycobacterial diseases, and ultimately may help to develop new therapeutic and/or prophylactic strategies against this insidious disease. □

Methods

Disruption of *hbhA* in *M. tuberculosis* and *M. bovis* BCG

A 4.1-kb DNA fragment encompassing the full-length *hbhA* gene¹⁶ was amplified by polymerase chain reaction (PCR) using 5'-CGTGCGGGAGAAATTCCTGTC-3' and 5'-ATATGCGGCGCCACCTCCAGCCTGACCAA-3' primers. The amplicon was digested by *NotI* and the resulting 3.5-kb fragment was cloned into the *NotI* site of pJQ200 (ref. 19), giving rise to pPK1. The *aph* cassette, conferring kanamycin resistance, was excised by *Bam*HI digestion from pUC4K (Amersham Pharmacia Biotech) and cloned into the unique *Bcl*I site located within the *hbhA* gene, leading to pPK2. pPK3 was obtained by cloning the 4.8-kb *NotI* fragment from pPK2, and containing the *hbhA::aph* construct, into *NotI* of pPR27 (ref. 4). *Mycobacterium bovis* BCG and *M. tuberculosis* 103 were

electroporated (2.5 kV, 800 Ω , 25 μ F) with 2 μ g of pPK3 and subsequently plated onto 7H10 Middlebrook agar supplemented with 20 μ g ml⁻¹ kanamycin (Sigma) and incubated at 32 °C. Several kanamycin-resistant clones were grown for 2 weeks in 7H9 Middlebrook medium supplemented with 20 μ g ml⁻¹ kanamycin, and then plated onto 7H10 Middlebrook agar containing 20 μ g ml⁻¹ kanamycin and 2% sucrose. The kanamycin-resistant clones, which were no longer sensitive to sucrose, were screened by PCR using 5'-ATGACATCAA GGCTCCGTTGCTT-3' and 5'-TCTGCGATGCGACCGTACCCA-3' oligonucleotides flanking the *Bcl*I site in *hbhA*. Clones for which a 1.8-kb fragment was amplified were analysed by Southern blotting.

Complementation of *hbhA::aph* strains

The 0.9-kb *Sa*I fragment containing the full-length *hbhA* gene with its promoter¹⁶ was treated with T4 DNA polymerase (Roche) to create blunt ends, and subsequently cloned into the *Eco*RV site of pYUB415 (ref. 17), giving rise to pPK4. *hbhA::aph* strains were electroporated with 2 μ g pPK4, followed by plating onto 7H10 Middlebrook agar supplemented with 50 μ g ml⁻¹ hygromycin (Roche) and 20 μ g ml⁻¹ kanamycin. We tested hygromycin-resistant clones for HBHA production by immunoblotting.

Interaction of *hbhA::aph* strains with eukaryotic cells

Cytoadherence assays were performed using the murine J774.1 macrophage-like cells (ATCC, TIB67), murine MH-S alveolar macrophage-like cells (ATCC, CRL 2019), human U937 macrophage-like cells (ATCC, CRE 1592.2) or human A549 type II pneumocytes (ATCC, CCL 185) as described¹³, except that mycobacteria and target cells were incubated for 1 h using a multiplicity of infection (MOI) of ten. Cell monolayers were then lysed with 0.1% Triton X-100, and serial dilutions were plated onto 7H11 medium for colony counting. We carried out invasion and intracellular survival assays as described²⁰.

Growth and persistence of *hbhA::aph* strains

We maintained mice according to the Institut Pasteur de Lille guidelines for laboratory animal husbandry. Eight-week-old female BALB/c mice were infected by the intranasal route as described¹⁰ with 10^3 and 10^6 CFUs of parental, mutant or complemented *M. tuberculosis* 103 and *M. bovis* BCG, respectively. Intravenous infections were performed in the tail vein with 10^6 viable units of parental, mutant or complemented strains. At indicated times, 4 mice per group were killed, and individual lungs and spleens were removed and homogenized. Serial dilutions were plated onto Middlebrook 7H10 agar for colony counting. The *in vivo* stability of pYUB415 in *hbhA*-complemented strains was controlled at each time point by plating simultaneously onto Middlebrook 7H10 agar with or without 50 μ g ml⁻¹ hygromycin.

To test the effects of pre-incubating mycobacteria with anti-HBHA antibodies on bacterial dissemination, 10^8 bacteria were incubated with 10 μ g of purified immunoglobulin for 2 h at 4 °C in a total volume of 500 μ l PBS, and 25 μ l (5×10^6 CFUs) were administered intranasally to each C57BL/6 mouse as described above. The source and purification of the monoclonal antibodies directed against HBHA, 4057 D2 and 3921 E4 as well as the control monoclonal antibody BPG10 have been described previously¹¹.

Received 8 March; accepted 18 May 2001.

- Dye, C., Scheele, S., Dolin, P., Pathania, V. & Ravigione, M. C. Consensus statement: global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *J. Am. Med. Assoc.* **282**, 677–686 (1999).
- Balasubramanian, V., Wiegand, E. H., Taylor, B. T. & Smith, D. W. Pathogenesis of tuberculosis: pathway to apical localization. *Tuber. Lung Dis.* **75**, 168–178 (1994).
- Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
- Pellicci, V. *et al.* Efficient allelic exchange and transposon mutagenesis in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **94**, 10955–10960 (1998).
- Berthet, F.-X. *et al.* Attenuation of virulence by disruption of the *Mycobacterium tuberculosis* *erp* gene. *Science* **282**, 759–762 (1998).
- McKinney, J. D. *et al.* Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature* **406**, 735–738 (2000).
- Camacho, L. R. *et al.* Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol. Microbiol.* **34**, 257–267 (1999).
- Cox, J. S., Chen, B., McNeil, M. & Jacobs, W. R. Jr Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* **402**, 79–83 (1999).
- Henderson, H. J., Dannenberg, A. M. Jr & Lurie, M. B. Phagocytosis of tubercle bacilli by rabbit pulmonary alveolar macrophages and its relation to native resistance to tuberculosis. *J. Immunol.* **90**, 553–556 (1963).
- Teitelbaum, R. *et al.* The M cells as a portal of entry to the lung for the bacterial pathogen *Mycobacterium tuberculosis*. *Immunity* **10**, 641–650 (1999).
- Menozi, F. D. *et al.* Identification of a heparin-binding hemagglutinin present in mycobacteria. *J. Exp. Med.* **184**, 993–1001 (1996).
- Reddy, V. M. & Kumar, B. Interaction of *Mycobacterium avium* complex with human respiratory epithelial cells. *J. Infect. Dis.* **181**, 1189–1193 (2000).
- Pethe, K. *et al.* *Mycobacterium smegmatis* laminin-binding glycoprotein shares epitopes with *Mycobacterium tuberculosis* heparin-binding haemagglutinin. *Mol. Microbiol.* **39**, 89–99 (2001).
- Pethe, K. *et al.* Characterization of the heparin-binding site of the mycobacterial heparin-binding hemagglutinin adhesin. *J. Biol. Chem.* **275**, 14273–14280 (2000).
- Delogu, G. & Brennan, M. J. Functional domains in the mycobacterial hemagglutinin, HBHA. *J. Bacteriol.* **181**, 7464–7469 (1999).
- Menozi, F. D. *et al.* Molecular characterization of the mycobacterial heparin-binding hemagglutinin, a mycobacterial adhesin. *Proc. Natl Acad. Sci. USA* **95**, 12625–12630 (1998).

17. Pavelka, M. S. & Jacobs, W. R. Jr Biosynthesis of diaminopimelate, the precursor of lysine and a component of peptidoglycan, is an essential function of *Mycobacterium tuberculosis*. *J. Bacteriol.* **178**, 6496–6507 (1996).
18. Teitelbaum, R. *et al.* A mAb recognizing a surface antigen of *Mycobacterium tuberculosis* enhances host survival. *Proc. Natl Acad. Sci. USA* **95**, 15688–15693 (1998).
19. Quandt, J. & Hynes, M. F. Versatile suicide vectors which allow direct selection for gene replacement in Gram-negative bacteria. *Gene* **127**, 15–21 (1993).
20. Wei, J. *et al.* Identification of a *Mycobacterium tuberculosis* gene that enhances mycobacterial survival in macrophages. *J. Bacteriol.* **182**, 377–384 (2000).

Acknowledgements

We thank W. R. Jacobs Jr for the gift of pYUB415; B. Gicquel for the gift of pPR27 and *M. tuberculosis* 103; J. P. Decavel for help in animal handling; and S. T. Cole, P. Bifani and A. R. Baulard for critically reading the manuscript. This work was supported by INSERM, Institut Pasteur de Lille, Région Nord-Pas de Calais, and the Ministère de la Recherche. K.P. and S.A. were supported by a fellowship from the Ministère de la Recherche and Aventis-Pasteur, respectively.

Correspondence and requests for materials should be addressed to C.L. (e-mail: camille.locht@pasteur-lille.fr).

Rab23 is an essential negative regulator of the mouse Sonic hedgehog signalling pathway

Jonathan T. Eggenschwiler, Edward Espinoza & Kathryn V. Anderson

Molecular Biology Program, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA

The mouse *open brain* (*opb*) and *Sonic hedgehog* (*Shh*) genes have opposing roles in neural patterning: *opb* is required for dorsal cell types and *Shh* is required for ventral cell types in the spinal cord^{1–3}. Here we show that *opb* acts downstream of *Shh*. Ventral cell types that are absent in *Shh* mutants, including the floor plate, are present in *Shh opb* double mutants. The organization of ventral cell types in *Shh opb* double mutants reveals that Shh-independent mechanisms can pattern the neural tube along its dorsal–ventral axis. We cloned *opb* by a map-based approach and found that it encodes Rab23, a member of the Rab family of vesicle transport proteins. The data indicate that dorsalizing signals activate transcription of *Rab23* in order to silence the Shh pathway in dorsal neural cells.

The extracellular signalling protein Sonic hedgehog (*Shh*) is essential for many aspects of mammalian embryogenesis, including patterning of the neural tube and limbs³. Inappropriate activation of the Hedgehog pathway after birth can promote development of tumours, including basal cell carcinoma and medulloblastoma^{4,5}. In *Drosophila*, Hedgehog acts as an extracellular ligand that binds the transmembrane receptor Patched⁶. In the absence of Hedgehog, Patched represses the activity of a second transmembrane protein, Smoothened, thereby blocking the downstream signalling pathway. The Hedgehog pathway acts through the transcription factor Ci⁷. Vertebrate homologues of Hedgehog, Patched, Smoothened and Ci (the Gli family) have been identified and seem to act in a signalling pathway similar to that characterized in *Drosophila*.

The mouse *open brain* (*opb*) gene was identified on the basis of two recessive mutations, one spontaneous and one induced by *N*-ethyl-*N*-nitrosourea (ENU)^{1,8}. Embryos homozygous for either mutant allele die during the second half of gestation, with an open neural tube in the head and spinal cord, abnormal somites, polydactyly, and poorly developed eyes. Mutations in *Shh* and *opb* cause opposing transformations in neural cell fate: *Shh* mutant embryos lack ventral cell types throughout the spinal cord, whereas

opb mutant embryos lack dorsal cell types specifically in the caudal spinal cord^{1–3}. The earliest requirement for *opb* is in the specification of dorsal neural progenitors, where *opb* acts in a cell autonomous manner². The *opb* mutant phenotypes resemble those produced by partial loss of function of *Patched* (ref. 9), the Shh receptor, which acts as a negative regulator of the Shh pathway. Thus the *opb* mutant phenotype could be the result of partial activation of the Shh signalling pathway in dorsal and lateral neural cells.

To investigate the relationship between *opb* and *Shh*, we examined the phenotype of embryos homozygous for *opb*² and a null mutation of *Shh*. *Shh opb* double mutant embryos were much larger than *Shh* mutant embryos and were similar in size to *opb* and wild-type littermates (Fig. 1a). *Shh* mutants are holoprosencephalic and lack ventral regions of the brain and head, whereas *opb* mutants are exencephalic. Ventral head structures were restored in *Shh opb* embryos and the double mutants showed the open brain and spinal cord characteristic of *opb* single mutants. The limbs of *Shh* mutants are truncated³, whereas the limbs of double mutants were of normal length and showed preaxial polydactyly similar to that of *opb* mutants. Thus many aspects of the morphology of the double mutant embryos were similar to that of *opb* single mutants, indicating that the *opb* mutation bypasses the requirement for *Shh* in several developmental contexts.

To analyse the double mutants at a molecular level, we compared expression of dorsal and ventral markers in the spinal cord in single and double mutants. Markers of the ventral cell types missing in *Shh* mutants were expressed in the spinal cord of *Shh opb* embryos including markers of motor neurons (Fig. 1b). In the double mutant, *Shh* was not required for specification of ventral cell types, suggesting that the Shh pathway is activated in the absence of ligand in *opb* mutants.

In both *Drosophila* and mice, transcription of *Patched* appears to be directly activated by Hedgehog signalling^{10,11}. We used a *Patched1-lacZ* reporter gene¹² to monitor *Patched1* expression in *opb* and *Shh opb* embryos. *Patched1-lacZ* expression was confined to the ventral half of the neural tube in wild type embryos at embryonic day 9.5 (E9.5). In contrast, *Patched1* expression expanded to include the entire caudal spinal cord of E9.5 *opb* and

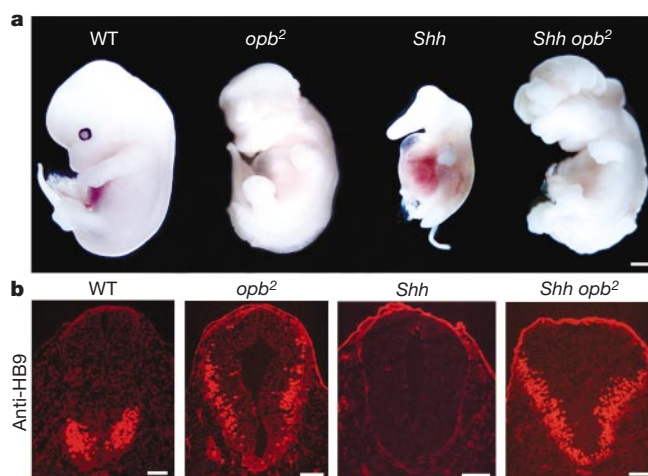


Figure 1 The *opb* mutation rescues the *Shh* mutant phenotype. **a**, External morphology of E13.5 embryos from *opb*² × *Shh* crosses. The *Shh opb*² mutant resembles the *opb*² single mutant—both fail to close the extreme rostral and caudal portions of the neural tube, lack eyes, exhibit polydactyly, and are normal in size. In contrast, the *Shh* mutant is one-third the size of wild type (WT), has truncated limbs and displays holoprosencephaly. **b**, Motor neuron precursor specification in lumbar regions revealed by anti-HB9 staining. Although motor neurons fail to be specified in *Shh* mutants, they are abundant in the ventral neural tube of *Shh opb*² double mutants. Similar results were obtained with the motor neuron markers *Isl1/2* and *Lim3* (data not shown). Scale bars: **a**, 1 mm; **b**, 50 μ m.

Shh opb mutant embryos (Fig. 2). These data show that *opb* acts as a direct negative regulator of the Shh signal transduction pathway.

The *Shh opb* neural tube showed a marked degree of patterning along the dorsal–ventral axis (Fig. 3). Similar results were observed with *Gli3*, a well established negative regulator of the Shh pathway¹³. In contrast to the *Shh Gli3* double mutant, in which some motor neurons were specified in caudal spinal cord, the complete range of ventral cell types was present in *Shh opb* double mutant embryos. Patterning in the caudal regions of the double mutant spinal cord was similar to that in *opb* single mutants (Fig. 3). As in *opb* single mutants, the dorsal marker Pax7 was not expressed and Pax6 was expressed in dorsal, rather than lateral, cells (Fig. 3; and data not shown). Nkx2.2, a marker of the V3 interneuron progenitors, was expressed in a lateral domain and excluded from the dorsal third of the neural tube in both *opb* single mutants and *Shh opb* double mutants. The floor plate marker Hnf3 β was expressed in *Shh opb* embryos in ventral neural cells, albeit in a narrower region than in the *opb* single mutants. Thus Shh ligand contributes to the activation of ventral target genes in the *opb* mutant neural tube. It is notable that the cells in the *Shh opb* double mutant embryos that expressed Hnf3 β had the characteristic morphology of floor plate cells, given that Shh can induce floor plate differentiation and *Shh* mutants do not make a floor plate^{3,14}.

In the double mutant embryos, the expression of *Patched-lacZ* was higher in ventral than in dorsal cells (Fig. 2b), indicating that a gradient of activity of the Shh pathway can be achieved in the absence of Shh itself. The identity of the signals that define this gradient of *Patched* activity and the asymmetry of neural cell types in the *Shh opb* double mutants is unclear. *Gli3* is expressed at high levels in the dorsal neural tube¹⁵ and should inhibit Shh pathway activity dorsally in the *Shh opb* double mutant. Bone morphogenetic protein (BMP) signals from the dorsal ectoderm and roof plate are also likely to have an important role¹⁶. However, it seems probable that another ventralizing signal from the notochord, in addition to Shh, is important in specifying extreme ventral cell fates, such as the floor plate. One candidate for a second ventralizing signal is the BMP inhibitor Noggin. Noggin is expressed in the notochord and is

required for the specification of ventral cell types¹⁷. Recent experiments have shown that BMPs antagonize Shh signalling, and have indicated that inhibition of BMP signalling is required to achieve full activation of the Shh pathway¹⁸. Thus, modulation of BMP

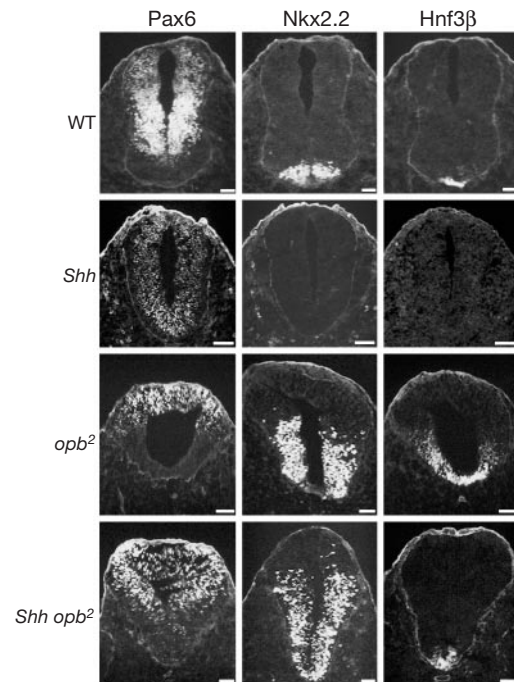


Figure 3 Neural patterning in *Shh opb*² mutants. Immunofluorescent staining for Pax6, Nkx2.2 and Hnf3 β in the lumbar regions of E10.5 embryos. Wild-type embryos express Pax6 in lateral cells; Nkx2.2 in ventral cells flanking the floor plate; and Hnf3 β in the floor plate. *Shh* mutants do not express markers of ventral fate (Nkx2.2 and Hnf3 β) and express low levels of Pax6 along the entire dorsal–ventral axis. In both *opb*² and *Shh opb*² double mutants, Pax6 is expressed at high levels in the dorsal neural tube, Nkx2.2 is expressed in the lateral and ventral regions, and Hnf3 β is expressed in cells at the ventral pole. Scale bars, 50 μ m.

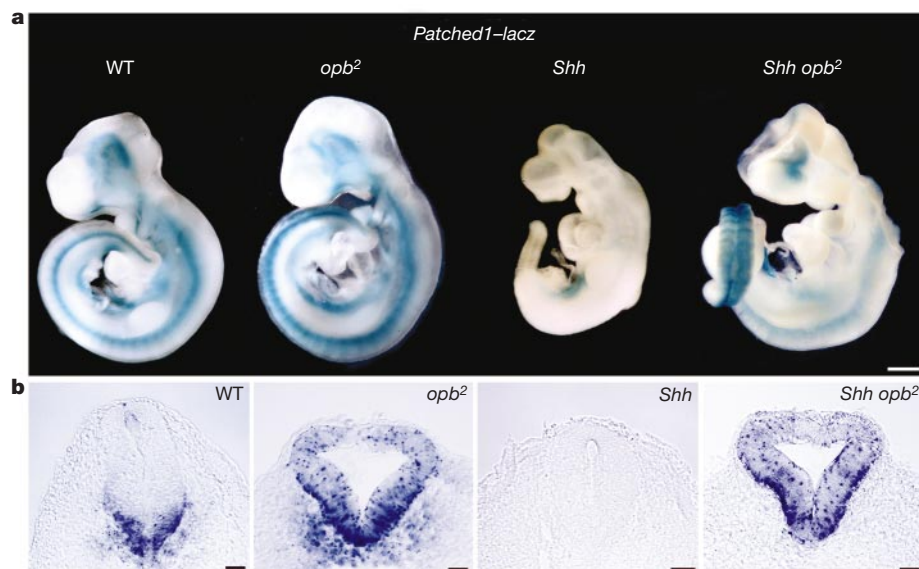


Figure 2 *Patched1-lacZ* expression in E9.5 embryos. Wild type, *opb*², *Shh* and *Shh opb*² mutant embryos heterozygous for a hypomorphic *Patched1-lacZ* allele. **a**, *Patched1-lacZ* expression revealed by X-gal staining is restricted to the ventral regions of the wild-type neural tube but is dorsally expanded in the *opb*² mutant. At this stage *Shh* expression is limited to the floor plate in *opb*² mutants². In the *Shh* mutant, *Patched1-lacZ* is only expressed in the gut, where it is presumably activated by local production of Indian

hedgehog. In contrast, *Shh*-independent expression of *Patched1-lacZ* is seen in the neural tube of the *Shh opb*² double mutant. **b**, Sections of embryos in **a** at lumbar levels. Graded *Patched1-lacZ* expression is expanded throughout the neural tubes of *opb*² and *Shh opb*² mutants relative to the wild type, but is not expressed in the *Shh* mutant neural tube. Scale bars: **a**, 500 μ m; **b**, 50 μ m.

signals may help produce a gradient of activity of the Shh signalling pathway, even in the absence of a gradient of Shh ligand.

We mapped the *opb* gene in interstrain and intersubspecific backcrosses of 1,435 animals to a 0.3 cM interval on proximal chromosome 1 between markers 13MMHAP88FRE7.seq and D1Mit318 (Fig. 4a). This interval was completely represented in two overlapping yeast artificial chromosomes (YACs). The genes in the mouse radiation hybrid map in the *opb* interval are homologous to genes on human chromosome 6p11–p12. Using the human genome sequence and mouse YACs, we found that three characterized genes lie in the *opb* interval: *DNA primase*, *Rab23* and *KIAA0576*. *DNA primase* was an unlikely candidate because DNA replication should be essential in all cells. *KIAA0576* seemed an unlikely candidate because it is not represented in embryonic complementary DNA libraries and the gene straddles the D1Mit318 marker that marks the limit of the *opb* interval.

Because *Rab23* is represented in embryonic cDNA libraries and lies entirely within the interval that includes the gene, we considered it as a candidate for *opb*. The *Rab23* gene encodes a member of the Rab family of GTPases that control vesicle transport, and is expressed at high levels in the adult brain¹⁹. We sequenced cDNAs

of the *Rab23* transcripts from *opb*¹ and *opb*² homozygous embryos and found single-nucleotide substitutions in both mutant alleles that introduced early stop codons into the open reading frame (Fig. 4b). Two independent recombinants (out of 1,435 opportunities) placed *opb*² proximal to D1Mit318. Both recombinant chromosomes (which lost the parental C57BL/6 allele of D1Mit318 but carried the *opb*² mutation) retained the stop codon mutation in *Rab23*, confirming that the stop codon substitution was never separated by recombination from the mutation responsible for the *opb* phenotype. The *opb*¹ and *opb*² alleles would produce truncated proteins of 39 and 79 amino acids, respectively, compared with the full-length product of 237 amino acids.

Rab23 is 30–35% identical to other mammalian Rab proteins and includes all the canonical motifs required for guanine nucleotide binding, GTP hydrolysis, membrane association and the conformational switch between the GTP- and GDP-bound state^{19–21}. *Rab23* is a relatively divergent Rab protein with an unusually long carboxy-terminal tail¹⁹. The gene most similar to the mouse and human *Rab23* genes is an uncharacterized *Drosophila* gene, which is 55% identical to mouse *Rab23* and seems to be a true homologue (Fig. 4b). The truncated proteins encoded by the *opb*¹ and *opb*²

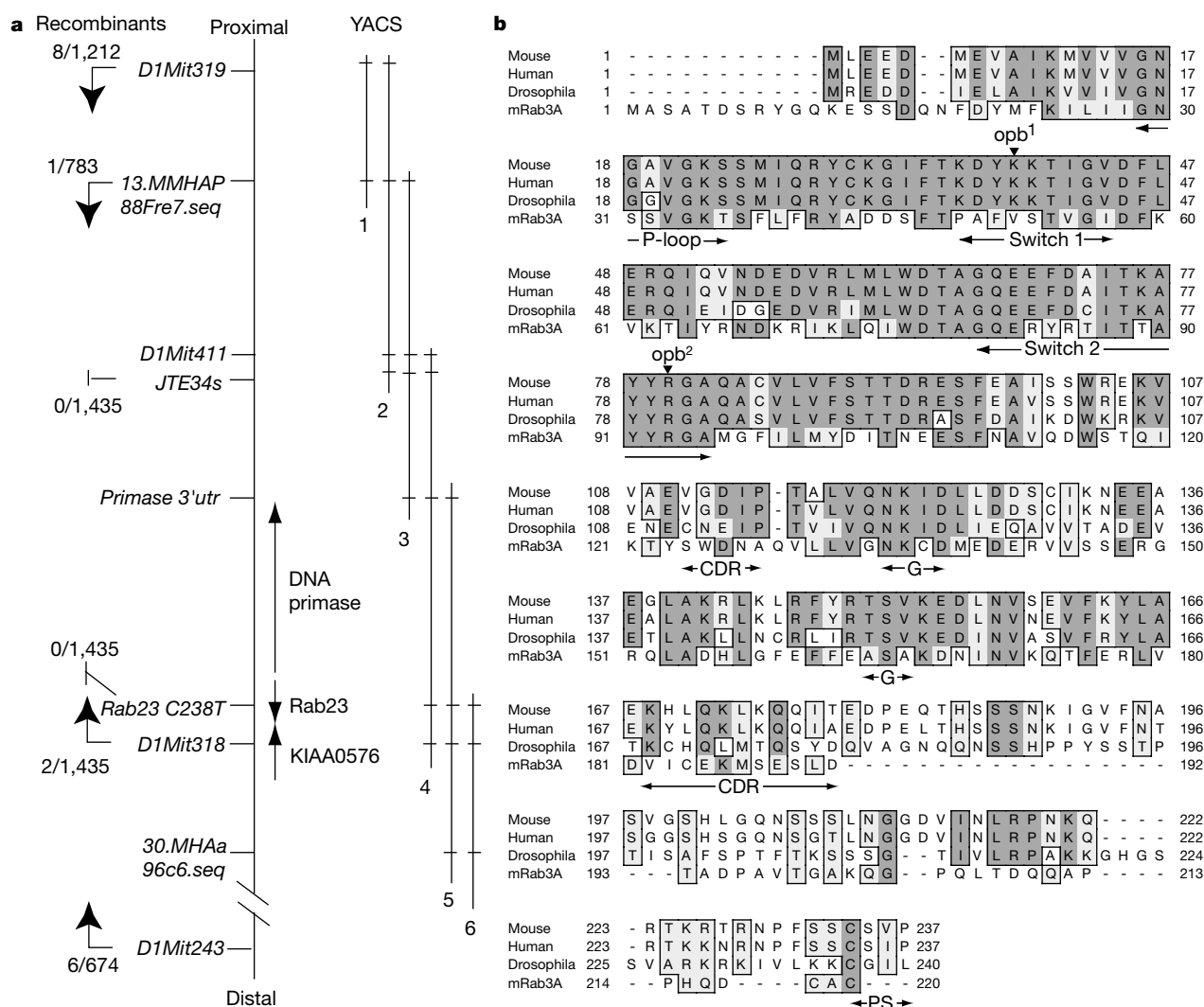


Figure 4 Molecular identification of *opb*. **a**, Genetic and physical mapping of the *opb* interval on proximal mouse chromosome 1. The number of recombination events/number of opportunities is indicated for each polymorphic marker. The marker JTE34s and the polymorphism generated by the C238T change in *Rab23* were never separated from *opb*. YAC clones 1–6 spanning the region correspond to WI/MIT-820 mouse clones 141B12, 332H9, 326B3, 343E11, 308G7 and 379F8, respectively. **b**, Alignment of mouse and

human *Rab23* amino-acid sequence with a related *Drosophila* gene (CG2108) and mouse *Rab3a*. Positions of stop codons in the *opb*¹ and *opb*² alleles are indicated above the sequence. Functional domains defined in other Rab family proteins are indicated below the sequence. P-loop, phosphate-binding loop; Switch 1 and 2, conformational switch regions; CDR, Rab effector-binding regions; G, guanine nucleotide-binding residues; PS, prenylation site.

alleles would lack the domains required for guanine nucleotide and Rab effector binding (Fig. 4b), and are therefore almost certainly null alleles.

The identification of a Rab protein as a specific component of the Shh pathway seems surprising, in part because vesicle transport is required in all cells. However, 60 mammalian *Rab* genes have been identified^{20,22}, and many Rab proteins are expressed in specific cell types and are localized to specific membrane compartments^{23,24}. Because all the *opb* phenotypes are also caused by mutations in genes such as *Patched1* and *Gli3*, which antagonize the Shh signalling pathway^{9,25}, *Rab23* may be dedicated to the regulation of Hedgehog signalling.

In principle, a Rab protein could act in the Shh pathway by controlling the secretion of specific signalling components or transcytosis of Shh ligand. However, chimaera experiments showed that the primary defect in *opb* mutants is cell autonomous². In addition, the double mutants show that *opb* bypasses the requirement for *Shh*, and so *opb* does not control secretion or transport of Shh. Therefore, *opb* must act intracellularly to control some aspect of the Shh signalling pathway.

Previous results had suggested that vesicle trafficking could play a

role in Hedgehog signal transduction. Most Patched protein is present in intracellular vesicles^{26,27}, and both Patched and Smoothed appear to shuttle between the plasma membrane and cytoplasmic vesicles in response to Hedgehog ligand²⁸. In addition, Patched is similar in structure to the Niemann Pick C1 (NPC1) protein²⁹, which promotes the movement of cholesterol and glycolipids from lysosomes to the plasma membrane, a process that appears to require Rab activity³⁰.

At E10.5, the *Rab23* RNA was present at low levels in most tissues, and was present at high levels in the spinal cord, somites, limb buds and cranial mesenchyme (Fig. 5a). In the spinal cord, *Rab23* was expressed at highest levels in the dorsal half of the neural tube, although it was excluded from the roof plate (Fig. 5c). In the limb bud, it was expressed in the crescent of mesenchymal cells that are capable of responding to Shh signalling. The expression pattern of *Rab23* RNA was similar to that of *Gli3*, another negative regulator of the Shh signalling pathway—*Gli3* is expressed prominently in the dorsal half of the spinal cord, somites, the limb buds and cranial mesenchyme¹⁵.

Rab23 expression was restricted to dorsal regions of the *Shh* mutant neural tube in a pattern similar to wild type (Fig. 5d), indicating that ventral repression by Shh does not define the *Rab23* expression domain. In contrast, *Rab23* expression was not detected in the caudal neural tube of *opb* mutants (Fig. 5f), although expression in the rostral neural tube was normal (Fig. 5b, e). One explanation of the results is that the signals that activate the expression of *Rab23* in the dorsal neural tube originate in the dorsal cells, such as roof plate, that fail to be specified in the caudal neural tube of the *opb* mutant. Signals from the roof plate, such as BMPs, are important for the specification of dorsal neural cell types¹⁶. Thus, silencing of the Shh pathway through activation of *Rab23* expression may be a part of the mechanism by which BMPs promote dorsal neural cell fates. □

Methods

Mouse strains

The *opb*¹ (ref. 1) and *opb*² (refs 2, 8) mutations were studied on a C3H background; both mutants produced indistinguishable phenotypes when compared in the same genetic background². We genotyped mice and embryos using polymorphic PCR markers^{2,8}. *Shh* mice and embryos were genotyped as described³. *opb*²;*Shh*^{+/−} mice were intercrossed to yield *opb*²/*opb*²;*Shh*^{+/−} double mutants, which were analysed in parallel with their wild type and single mutant littermates. Neither *opb*² nor *Shh* heterozygosity influenced the phenotype in any mutant combination. *Patched1*–*lacZ* (Ptc108D11) is a knock-in hypomorphic allele that expresses *lacZ* in a pattern that faithfully mimics *Patched1* expression¹² (M. P. Scott, personal communication).

Immunohistochemistry and *in situ* hybridization

Immunofluorescent staining of fixed, frozen sections, X-gal staining for β -galactosidase activity in whole-mount embryos, and *in situ* hybridization to whole-mount embryos were carried out using antibodies and protocols as described².

Genetic and physical mapping of the *opb* locus

Genetic mapping was performed using the *opb*² allele in C3H/HeJ and MOLF/Ei (*Mus musculus molossinus*) backcrosses for a total of 652 and 783 informative opportunities for recombination, respectively. *opb*² mapped to a 0.3 cM interval between markers 13MMHAP88FRE7.seq and D1Mit318 (at map position 18.5). It failed to segregate from the markers D1Mit411 (783 opportunities), JTE34s (1,435 opportunities) and *Rab23C238T* (1,435 opportunities). Clones from the WI/MIT820 Mouse YAC library (Research Genetics) and from the RPCI-22 and RPCI-23 Mouse BAC libraries (CHORI) were identified by database searches (MGI, <http://www.informatics.jax.org>) and library filter hybridization with labelled oligonucleotides and PCR products. Clones were anchored and ordered by sequence-tagged site (STS) content mapping. Additional STS makers were generated from BAC end sequence obtained from the BES database (http://www.tigr.org/tldb/humgen/bac_end_search/bac_end_search.html) and by direct sequencing of BAC DNA using T7 and SP6 primers. Additional information on markers and mapping experiments are available on request. Synteny with the human genome was established using mouse RH data (<http://www.informatics.jax.org>) in conjunction with our physical mapping data and human draft genomic sequence (<http://genome.cse.ucsc.edu>).

Molecular biology

The *Rab23* coding region (plus portions of the 5' and 3' untranslated regions) was

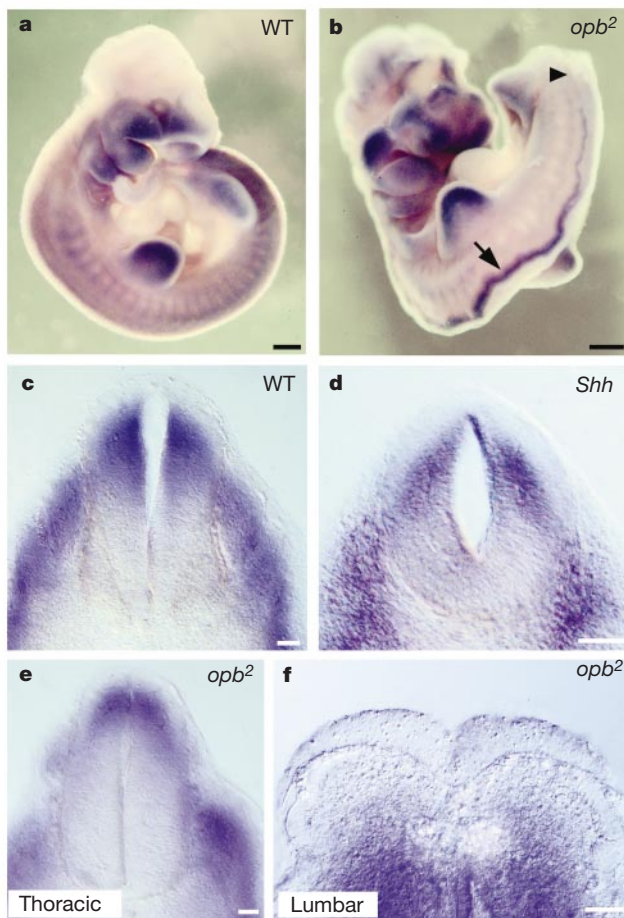


Figure 5 *Rab23* expression in E10.5 embryos. **a**, In wild type, *Rab23* is expressed at high levels in the neural tube, limb mesoderm, branchial arches, eye and somites. **b**, *Rab23* is expressed normally in most tissues of the *opb*² mutant. However, neural tube expression, which appears roughly normal at the thoracic level (arrow), is absent in the open, posterior neural tube (arrowhead). **c–f**, Sections through thoracic (**c**, **d**, **e**) and lumbar levels (**f**) of wild type (**c**) *Shh* mutant (**d**) and *opb*² mutant (**e**, **f**) embryos. *Rab23* is expressed in a graded manner in the dorsal half of the wild-type neural tube (**c**) but is not expressed in the roof plate. The expression pattern in the neural tube of the *Shh* mutants (**d**) is similar. In the *opb*² neural tube, the *Rab23* expression domain is normal rostrally, slightly narrowed in the thoracic region (**e**), and is absent in lumbar region (**f**). The neural tube in **f** failed to close as a result of the *opb* mutant phenotype. Scale bars: **a**, **b**, 700 μ m; **c–f**, 50 μ m).

amplified by RT-PCR (One step, Life Technologies) from E10.5 *opb¹*, *opb²* and C57BL/6J embryos using forward and reverse primers (5'-TGGTTT GAGGAGACAGTGTGG-3' and 5'-CGCCATCTACATTCTGAGGAGCAC-3', respectively). cDNAs from three embryos per genotype were separately amplified, subcloned and sequenced, yielding consistent results. An A to T change at position +115 and a C to T change at position +238 of the open reading frame (both causing nonsense mutations) were observed in *opb¹* and *opb²* mutants, respectively. The change in *opb²* created a Taq1 restriction-fragment length polymorphism, which was assayed in genomic DNA using the marker *Rab23C238T* (forward: 5'-ATGCCCTCTGTGATGGAGACTG-3'; reverse: 5'-AACACCCCGTCATT CAAAC-3') to confirm the linkage of *opb²* with the change at position +238. Genotyping with *Rab23C238T* confirmed the change in the genomic DNA of eight additional *opb²* homozygous embryos.

Received 21 January; accepted 1 June 2001.

- Günther, T., Struwe, M., Aguzzi, A. & Schughart, K. *Open brain*, a new mouse mutant with severe neural tube defects, shows altered gene expression patterns in the developing spinal cord. *Development* **120**, 3119–3130 (1994).
- Eggenschwiler, J. T. & Anderson, K. V. Dorsal and lateral fates in the mouse neural tube require the cell-autonomous activity of the *open brain* gene. *Dev. Biol.* **227**, 648–660 (2000).
- Chiang, C. *et al.* Cyclopia and defective axial patterning in mice lacking *Sonic hedgehog* gene function. *Nature* **383**, 407–413 (1996).
- Goodrich, L. V. & Scott, M. P. Hedgehog and patched in neural development and disease. *Neuron* **21**, 1243–1257 (1998).
- Matise, M. P. & Joyner, A. L. *Gli* genes in development and cancer. *Oncogene* **18**, 7852–7859 (1999).
- Ingham, P. W. Transducing Hedgehog: the story so far. *EMBO J.* **17**, 3505–3511 (1998).
- Method, N. & Basler, K. An absolute requirement for *Cubitus interruptus* in Hedgehog signaling. *Development* **128**, 733–742 (2001).
- Kasarskis, A., Manova, K. & Anderson, K. V. A phenotype-based screen for embryonic lethal mutations in the mouse. *Proc. Natl Acad. Sci. USA* **95**, 7485–7490 (1998).
- Milenkovic, L., Goodrich, L. V., Higgins, K. M. & Scott, M. P. Mouse *patched1* controls body size determination and limb patterning. *Development* **126**, 4431–4440 (1999).
- Forbes, A. J., Nakano, Y., Taylor, A. M. & Ingham, P. W. Genetic analysis of hedgehog signalling in the *Drosophila* embryo. *Development* (Suppl.) **119**, 115–124 (1993).
- Goodrich, L. V., Johnson, R. L., Milenkovic, L., McMahon, J. A. & Scott, M. P. Conservation of the hedgehog/patched signaling pathway from flies to mice: induction of a mouse *patched* gene by Hedgehog. *Genes Dev.* **10**, 301–312 (1996).
- Goodrich, L. V., Milenkovic, L., Higgins, K. M. & Scott, M. P. Altered neural cell fates and medulloblastoma in mouse *patched* mutants. *Science* **277**, 1109–1113 (1997).
- Litingtung, Y. & Chiang, C. Specification of ventral neuron types is mediated by an antagonistic interaction between *shh* and *gli3*. *Nature Neurosci.* **10**, 979–985 (2000).
- Echelard, Y. *et al.* Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity. *Cell* **75**, 1417–1430 (1993).
- Mo, R. *et al.* Specific and redundant functions of *Gli2* and *Gli3* zinc finger genes in skeletal patterning and development. *Development* **124**, 113–23 (1997).
- Lee, K. J., Dietrich, P. & Jessell, T. M. Genetic ablation reveals that the roof plate is essential for dorsal interneuron specification. *Nature* **403**, 734–740 (2000).
- McMahon, J. A. *et al.* Noggin-mediated antagonism of BMP signaling is required for growth and patterning of the neural tube and somite. *Genes Dev.* **12**, 1438–1452 (1998).
- Liem, K. F., Jessell, T. M. & Briscoe, J. Regulation of the neural patterning activity of sonic hedgehog by secreted BMP inhibitors expressed by notochord and somites. *Development* **127**, 4855–4866 (2000).
- Olkkonen, V. M. *et al.* Isolation of a mouse cDNA encoding Rab23, a small novel GTPase expressed predominantly in the brain. *Gene* **138**, 207–211 (1994).
- Perreira-Leal, J. B. & Seabra, M. C. The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily. *J. Mol. Biol.* **301**, 1077–1087 (2000).
- Ostermeier, C. & Brunker, A. T. Structural basis of Rab effector specificity: crystal structure of the small G protein Rab3A complexed with the effector domain of rabphilin-3A. *Cell* **96**, 363–374 (1999).
- Bock, J. B., Matern, H. T., Peden, A. A. & Scheller, R. H. A genomic perspective on membrane compartment organization. *Nature* **409**, 839–841 (2001).
- Sonnichsen, B., De Renzis, S., Nielsen, E., Rietdorf, J. & Zerial, M. Distinct membrane domains on endosomes in the recycling pathway visualized by multicolor imaging of Rab4, Rab5, and Rab11. *J. Cell Biol.* **149**, 901–914 (2000).
- Wilson, S. M. *et al.* A mutation in Rab27a causes the vesicle transport defects observed in ashen mice. *Proc. Natl Acad. Sci. USA* **97**, 7933–7938 (2000).
- Hui, C. C. & Joyner, A. L. A mouse model of Greig cephalopolysyndactyly syndrome: the *Extra-toes* mutation contains an intragenic deletion of the *Gli3* gene. *Nature Genet.* **3**, 241–246 (1993).
- Capdevila, J., Pariente, E., Sampedro, J., Alonso, J. L. & Guerrero, I. Subcellular localization of the segment polarity protein *patched* suggests an interaction with the wingless reception complex in *Drosophila* embryos. *Development* **4**, 987–998 (1994).
- Incardona, J. P. *et al.* Receptor-mediated endocytosis of soluble and membrane-tethered Sonic hedgehog by Patched-1. *Proc. Natl Acad. Sci. USA* **97**, 12044–12049 (2000).
- Denef, N., Neubuser, D., Perez, L. & Cohen, S. M. Hedgehog induces opposite changes in turnover and subcellular localization of patched and smoothened. *Cell* **120**, 521–531 (2000).
- Carstea, E. D. *Niemann-Pick C1* disease gene: homology to mediators of cholesterol homeostasis. *Science* **277**, 228–231 (1997).
- Holttu-Vuori, M., Maatta, J., Ullrich, O., Kuusmanen, E. & Ikonen, E. Mobilization of late-endosomal cholesterol is inhibited by Rab guanine nucleotide dissociation inhibitor. *Curr. Biol.* **10**, 95–98 (2000).

Acknowledgements

Monoclonal antibodies were obtained from the Developmental Studies Hybridoma Bank developed under the auspices of the NICHD and maintained by The University of Iowa, Department of Biological Sciences. We thank P. Beachy and M. Scott for discussions and for gifts of the *Shh* mice and *Patched-lacZ* mice; M. Rosen for discussions about Rab structure; and M. J. García-García, K. Brennan, J. Timmer, L. Niswander and T. Bestor for

helpful comments on the manuscript. This work was supported by an NIH grant (K.V.A.), the Lita Annenberg Hazen Foundation and an ACS postdoctoral fellowship to J.T.E.

Correspondence and requests for materials should be addressed to K.V.A. (e-mail: k-anderson@ski.mskcc.org).

Phagocytosis promotes programmed cell death in *C. elegans*

Peter W. Reddien*, Scott Cameron*† & H. Robert Horvitz*

*Howard Hughes Medical Institute, Department of Biology, 68-425, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

†Division of Pediatric Hematology Oncology, Children's Hospital/Dana Farber Cancer Institute, Boston, Massachusetts 02115, USA

In the nematode *Caenorhabditis elegans* programmed cell death requires the killer genes *egl-1*, *ced-4* and *ced-3* (refs 1 and 2), and the engulfment of dying cells requires the genes *ced-1*, *ced-2*, *ced-5*, *ced-6*, *ced-7*, *ced-10* and *ced-12* (refs 3–5). Here we show that engulfment promotes programmed cell death. Mutations that cause partial loss of function of killer genes allow the survival of some cells that are programmed to die, and mutations in engulfment genes enhance the frequency of this cell survival. Furthermore, mutations in engulfment genes alone allow the survival and differentiation of some cells that would normally die. Engulfment genes probably act in engulfing cells to promote death, as the expression in engulfing cells of *ced-1*, which encodes a receptor that recognizes cell corpses⁶, rescues the cell-killing defects of *ced-1* mutants. We propose that engulfing cells act to ensure that cells triggered to undergo programmed cell death by the CED-3 caspase⁷ die rather than recover after the initial stages of death.

During metazoan development, programmed cell death acts in the shaping of tissues, the refinement of neuronal connections, and the removal of unnecessary or damaged cells⁸. Cells dying by programmed cell death are recognized and rapidly engulfed by phagocytic cells⁹. In *C. elegans*, engulfment can begin even before the completion of the cell division that generates a cell programmed to die¹⁰. *C. elegans* genes that control the killing and engulfment processes of programmed cell death have been identified, and the molecular mechanisms underlying these processes in *C. elegans* have proven to be evolutionarily conserved^{11,12}. On the basis of their genetic interactions, the engulfment genes fall into two partially redundant sets, *ced-1*, *ced-6*, *ced-7* and *ced-2*, *ced-5*, *ced-10* and *ced-12*, which have been proposed to define two parallel and partially redundant pathways (ref. 4; and Z. Zhou and H.R.H., unpublished results). The CED-7 ABC transporter¹³ functions with the CED-1 SREC receptor in corpse recognition⁶. CED-6 is an adaptor-like protein¹⁴ and may transduce signals from CED-1. The CED-2 CrkII, CED-5 DOCK180, and CED-10 Rac proteins have been proposed to act in a signalling pathway that controls cytoskeletal extensions of cell surfaces around dying cells during engulfment^{15,16}. That cell corpses are formed in animals which have defective cell-corpse engulfment indicates that engulfment is not essential for cells to die by programmed cell death.

We sought new killer genes by screening for mutations (P.W.R. and H.R.H., unpublished results) that enhanced the defect in programmed cell death of the *ced-3* partial-loss-of-function mutation *n2427* (ref. 17). Unexpectedly, we isolated mutations in engulfment genes. We assayed defects in programmed cell death by counting the number of extra cell nuclei ('undead cells') in the anterior pharynxes of *C. elegans*, in which 16 cell deaths normally

occur during embryogenesis¹⁷. Animals carrying *ced-3(n2427)* had an average of 1.5 extra cells in their anterior pharynges, whereas double mutants carrying *ced-3(n2427)* and mutations in engulfment genes had a greater number of extra cells in this region: 4.5–6.2 (Table 1, $P < 0.0001$, unpaired *t*-test). Mutations in the cell-corpse engulfment genes alone did not result in extra cells in the anterior pharynx (Table 1). The effects of engulfment mutations on cell killing were not limited to the pharynx, because double mutant animals carrying both *ced-3(n2427)* and an engulfment gene mutation displayed a substantially increased frequency of survival of V5R.paapp, a cell that dies postembryonically in the lateral ectoderm⁹, compared with that of the single mutants (Table 1). For example, V5R.paapp never survived in *ced-3(n2427)* or *ced-6(n1813)* animals but survived in 40% of

ced-6(n1813); ced-3(n2427) animals.

The engulfment mutation *ced-1(e1735)* enhanced multiple *ced-3* alleles, *ced-4* alleles (for example, *ced-4(n3312)* for V5R.paapp, $P < 0.0001$, unpaired *t*-test), and an *egl-1* allele ($P < 0.0001$, unpaired *t*-test) (Table 1). Mutations in all known engulfment genes enhanced *ced-3(n2427)* (Table 1). The engulfment genes *ced-1*, *ced-2*, *ced-5*, *ced-6* and *ced-10* appear to act in engulfing cells for cell-corpse phagocytosis^{6,14–16}, and *ced-7* appears to act in both engulfing and dying cells¹³. That mutations in all engulfment genes, including those of both pathways, enhanced the cell-death defects caused by *ced-3(n2427)* suggests that some aspect of the engulfment process itself is involved in the killing of cells.

We also examined the effects of engulfment gene mutations on cell death in the ventral nerve cord. Postembryonic development in

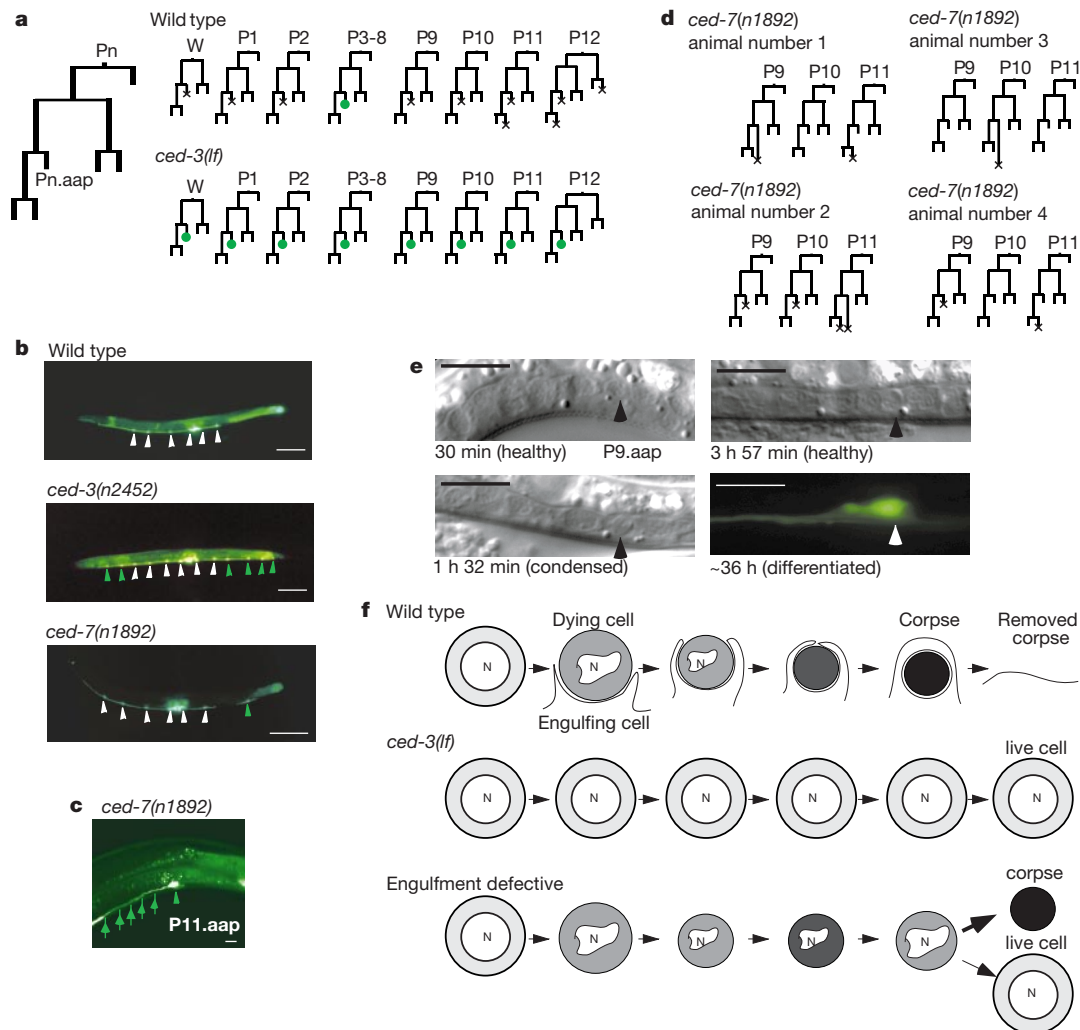


Figure 1 Engulfment mutations lead to the survival of cells programmed to die in the ventral cord. **a**, 13 blast cells (W and P1–P2) divide post-embryonically to generate motor neurons of the ventral cord⁹. P3.aap–P8.aap (the posterior daughter of the anterior daughter of the anterior daughter of the P3–P8 blast cells) differentiate into VC motor neurons and express GFP from the *lin-11* promoter. W.ap, P1.aap, P2.aap and P9.aap–P12.aap undergo programmed cell death. In a *ced-3* mutant, W.ap, P1.aap, P2.aap and P9–P12.aap survive and differentiate into VC-like neurons and express GFP from the *lin-11* promoter (see **b**). **b**, A wild-type adult displayed fluorescence from P3.aap–P6.aap, vulval tissue, and P7.aap–P8.aap. A *ced-3(n2452)* adult additionally displayed fluorescence from the undead P1.aap, P2.aap, and P9.aap–P12.aap cells. Expression from undead W.ap and P1.aap is variable. A *ced-7(n1892)* adult displayed fluorescence in P11.aap. White arrowheads, P3.aap–P8.aap; green arrowheads, Pn.aap cells that failed to die. Anterior, left. Ventral, down. Scale bar, 100 μ m. **c**, P11.aap in a *ced-7(n1892)* animal survived, expressed *P_{lin-11}gfp*, and extended an axon. Arrowhead, cell body. Green

arrows follow the P11.aap axon trajectory. Anterior, left. Ventral, down. Scale bar, 10 μ m. **d**, Cell lineages of *ced-7(n1892)* animals had normal division patterns but abnormalities in cell death. The deaths of P9.aap in animal number 1, P11.aap in animal number 2, and P10.aap in animal number 3 were delayed. In cases in which a cell lineally fated to die failed to do so, animals were observed for at least three hours after the last cell division occurred. **e**, A P9.aap cell in a *ced-7(n1892)* animal showed morphological characteristics of cell death 1 h 32 min after the cell was born. This cell recovered a normal nuclear morphology by 4 h and 36 h later, expressed GFP from the *lin-11* promoter. Black arrowhead, P9.aap. Anterior, left. Ventral, down. Scale bar, 10 μ m. **f**, Model for the effects of engulfment defects on cell killing. In *ced-3(lf)* animals cell death is not initiated. Engulfment is necessary for the efficient execution of death and in engulfment-defective animals cell death is initiated but cells typically attempt to recover and occasionally survive. N, nucleus.

the ventral cord involves 13 neuroblasts, termed W, and P1–P12, which together generate 10 cells that undergo programmed cell death⁹. Six of the ventral-cord cells that die are Pn.aap cells (posterior daughters of the anterior daughters of the anterior daughters of P blast cells), specifically, those generated by the P1, P2 and P9–P12 lineages; by contrast, the Pn.aap cells generated by the P3–P8 lineages survive and differentiate into class VC motor neurons⁹ (Fig. 1a). We found that the gene *lin-11*, which encodes a LIM-domain transcription factor¹⁸ and is expressed in VC motor neurons (G. Acton, S.C. and H.R.H. unpublished observations), was also expressed in the undead VC-equivalent cells (W.a.p, P1.aap, P2.aap and P9–12.aap) of a *ced-3* mutant (Fig. 1a, b), suggesting that these undead cells adopted a VC-like fate. We used the *P_{lin-11}gfp* reporter *nIs96* (see Methods) to assay the effects of engulfment mutations on Pn.aap cell deaths.

Mutations in each of the engulfment genes resulted in the low-penetrance survival of Pn.aap cells that normally die (Table 2). (That the *lin-11* promoter does not express until the L4 larval stage (our unpublished observations) and the deaths of the Pn.aap cells occur approximately 24 h earlier in the late L1/early L2 larval stage indicated that the ventral-cord fluorescence we observed was the result of transcription and translation in an undead VC-like cell rather than the persistence of green fluorescent protein (GFP) in a cell corpse.) Pn.aap cells that failed to die in engulfment mutants showed characteristics of differentiated VC neurons: neuronal nuclear morphology, elongated axons and expression of *lin-11* (Fig. 1c). We confirmed that the fluorescent cells we observed were undead VC-like neurons by examining the patterns of ventral-cord corpses in *ced-1(e1735)* L2 larvae using Nomarski

microscopy, allowing these animals to develop into adults, and then determining the subsequent patterns of *P_{lin-11}gfp*-expressing cells (data not shown).

To determine if the undead VC-like cells in engulfment mutants expressed any morphological signs of initiating programmed cell death (and also to confirm directly that the ectopic fluorescent cells were undead Pn.aap cells), we directly observed the P cell lineages of *ced-7(n1892)* mutants. Although the P cell division patterns were normal, there were abnormalities in programmed cell deaths (Fig. 1d). In many cases the cells that failed to die initially displayed some of the morphological characteristics of dying cells, that is, cytoplasmic and nuclear condensation and loss of nuclear refractility. These morphological features were episodic and fluctuated between condensed and non-condensed states, eventually resulting in either a delay in the formation of a normal cell corpse or in the ultimate appearance of a normal-looking VC neuron (Fig. 1e). We observed similar episodic morphological changes in *ced-7(n1996)* and *ced-6(n1813)* mutants (data not shown). As cells that failed to die initially displayed some characteristics of dying cells, we suggest that engulfment is not necessary for the initiation of cell death but rather for the efficient execution of death once it has been initiated. We propose that engulfment actively promotes the killing process rather than passively eliminates the opportunity for recovery. In the latter case, we might expect that in the absence of engulfment the cell-death process would proceed normally up to a certain stage, but that the cells occasionally would subsequently recover. By contrast, we observed that in the absence of engulfment, cells that normally undergo programmed cell death instead often underwent episodic changes in morphology and sometimes appeared to fail to express any sign whatsoever of normal cell-corpse morphology, indicating that the cell-death process itself was abnormal in these cells. For example, two cells indicated in Fig. 1d (P11.aap in *ced-7(n1892)* animal number 3 and P10.aap in animal number 4) showed no or minor characteristics of dying cells. CED-3 caspase activity is necessary for the initiation of the morphological changes we observed, because these changes, which appear to reflect attempts at cell death, did not occur in *ced-3(n2452)* ventral-cord lineages (data not shown). (*n2452* is a *ced-3* allele deleted for the protease-

Table 1 Mutations in engulfment genes enhance cell survival

Genotype	Percentage of animals with undead V5R.paapp cells*	Number of extra cells (anterior pharynx)†
N2	0	0.1 ± 0.05
<i>ced-3(n2452)</i>	100	10.8 ± 0.32
<i>ced-1(e1735)</i>	4	0.1 ± 0.06
<i>ced-6(n1813)</i>	0	0.3 ± 0.09
<i>ced-7(n1892)</i>	8	0.1 ± 0.05
<i>ced-2(e1752)‡</i>	0	0.0 ± 0.03
<i>ced-5(n1812)</i>	0	0.0 ± 0.00
<i>ced-10(n1993)§</i>	4	ND
<i>ced-12(n3261)</i>	0	0.0 ± 0.0 (n = 25)
<i>ced-7(n1892); ced-5(n1812)</i>	12	ND
<i>ced-3(n2427)¶</i>	0	1.5 ± 0.14 (n = 47)
<i>ced-1(e1735); ced-3(n2427)¶</i>	48	5.9 ± 0.36
<i>ced-6(n1813); ced-3(n2427)¶</i>	40	5.9 ± 0.25
<i>ced-7(n1892); ced-3(n2427)¶</i>	56	6.2 ± 0.34
<i>ced-2(e1752); ced-3(n2427)¶</i>	32	4.5 ± 0.21
<i>ced-5(n1812); ced-3(n2427)¶</i>	36	5.0 ± 0.36
<i>ced-10(n1993); ced-3(n2427)¶</i>	44	ND
<i>ced-12(n3261); ced-3(n2427)¶</i>	40	5.6 ± 0.24 (n = 25)
<i>ced-7(n1892); ced-5(n1812); ced-3(n2427)¶</i>	84	ND
<i>ced-3(n2438)¶</i>	0	2.0 ± 0.20
<i>ced-1(e1735); ced-3(n2438)¶</i>	64	6.5 ± 0.31
<i>ced-3(n2447)</i>	15 (n = 27)	1.6 ± 0.14 (n = 70)
<i>ced-1(e1735); ced-3(n2447)</i>	57 (n = 30)	3.0 ± 0.27 (n = 36)
<i>ced-7(n1892); ced-3(n2447)</i>	80	4.4 ± 0.25 (n = 37)
<i>ced-3(n2443)</i>	8	1.9 ± 0.22
<i>ced-1(e1735); ced-3(n2443)</i>	20	3.5 ± 0.27
<i>ced-4(n3312)</i>	40 (n = 45)	2.1 ± 0.22
<i>ced-1(e1735); ced-4(n3312)</i>	80 (n = 46)	2.3 ± 0.22
<i>ced-4(n3195)¶</i>	23 (n = 30)	1.6 ± 0.24 (n = 25)
<i>ced-1(e1735); ced-4(n3195)¶</i>	80 (n = 30)	1.8 ± 0.36 (n = 25)
<i>ced-4(n3158)</i>	93 (n = 15)	3.7 ± 0.36 (n = 20)
<i>ced-1(e1735); ced-4(n3158)</i>	100 (n = 15)	4.6 ± 0.39 (n = 20)
<i>egl-1(n3331)#</i>	3 (n = 30)	3.7 ± 0.27 (n = 35)
<i>ced-1(e1735); egl-1(n3331)#</i>	47 (n = 30)	5.4 ± 0.33 (n = 35)

* The number of extra cells in the right postdeirid of L4 larvae was determined using Nomarski optics. In the wild type four cells are present, and one cell undergoes programmed cell death. Data are percentages. Sample size n = 25 unless otherwise shown.

† The number of extra cells in the anterior pharynxes of L3 larvae was determined using Nomarski optics. In the wild type, 16 cells undergo programmed cell death in this region. Data are means ± standard error of the means. Sample size n = 30 unless otherwise shown.

Strains contained the following additional mutations: ‡ *him-5(e1490)*; § *him-8(e1489)*; ¶ *unc-30(e191)*; ¶ *unc-69(e587)*; and # *nIs96*.

ND, not determined. Pharyngeal cell numbers could not be determined in these animals because of abnormal cell positions.

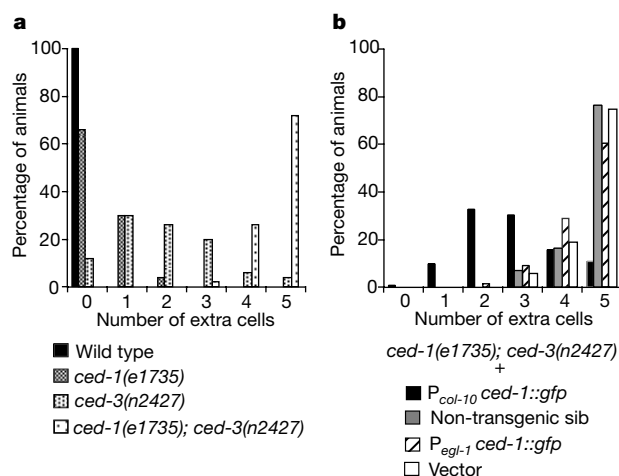


Figure 2 *ced-1* acts in engulfing cells to control cell killing. Distributions of the percentages of animals with specific numbers of extra cells in the ventral cord, as determined using *P_{lin-11}gfp* and observing fluorescent cell numbers in the regions where P2.aap and P9.aap–P12.aap reside. **a**, *ced-1(e1735)* caused some survival and enhanced the frequency of cell survival conferred by *ced-3(n2427)* in the ventral cord. 50 animals of each genotype were scored. **b**, *P_{col-10}ced-1::gfp* (engulfing cell expression) rescued the enhancement of *ced-3(n2427)* by *ced-1(e1735)* (n = 122). Non-transgenic siblings of transgenic *P_{col-10}ced-1::gfp* animals (n = 127), *P_{egl-1}ced-1::gfp* (dying cell expression) animals (n = 66), and *P_{col-10}gfp* (vector) animals (n = 79) showed distributions similar to that of *ced-1(e1735); ced-3(n2427)* animals in **a**.

encoding region of the *ced-3* gene¹⁷.) Perhaps if a dying cell fails to reach some critical point, the cellular alterations that result from CED-3 caspase activity can be reversed and cells can survive and differentiate (Fig. 1f).

Do engulfment genes promote death by acting in engulfing or in dying cells? Dying cells in the ventral cord are engulfed by hypodermal cells¹⁰, and expression of *ced-1* in the hypodermis can rescue *ced-1(lf)* engulfment defects⁶. *ced-1(e1735)* enhanced cell survival in the ventral cords of *ced-3(n2427)* animals (Fig. 2a), so we expressed *ced-1* in the hypodermis using the promoter of the *col-10* gene (V. Ambros, personal communication) (*P_{col-10}ced-1*; ref. 6) and in dying cells using the promoter of the *egl-1* gene¹⁹ (*P_{egl-1}ced-1*; ref. 6) in *ced-1(e1735); ced-3(n2427)* animals and determined the number of VC-like cells using the *P_{lin-11gfp}* reporter. Expression of *ced-1* in the hypodermis, but not in dying cells, rescued the enhancement of *ced-3(n2427)* conferred by *ced-1(e1735)* (Fig. 2b). We suggest that *ced-1*, and presumably other engulfment genes as well, acts within engulfing rather than within dying cells to promote death.

Does engulfment contribute to cell killing by inhibiting the protective function of the anti-apoptotic gene *ced-9*, which encodes a protein similar to the human anti-apoptotic protein Bcl-2 (ref. 20)? We tested whether *ced-1* mutations could enhance *ced-3(n2427)* in the absence of *ced-9* function. Defects in engulfment can significantly decrease killing even in the absence of *ced-9* function: *ced-9(n2812); ced-3(n2427)* animals had 6.3 extra cells in their pharynges, whereas *ced-1(e1735); ced-9(n2812); ced-3(n2427)* animals had 9.8 extra cells (Table 3) ($P < 0.0001$, unpaired *t*-test). Both *ced-9(n2812)* and *ced-1(e1735)* are probably null alleles^{6,20}. This observation suggests that engulfment does not kill cells by inhibiting the protective function of *ced-9* but rather acts downstream of or parallel to *ced-9* protection. In addition to its protective function *ced-9* also has a killing function: whereas *ced-9* loss-of-function mutations result in the deaths of cells that should live, these mutations also enhance the cell-killing defects conferred by partial *ced-3* loss-of-function mutations²¹. The promotion of programmed cell death by engulfment appears to occur by a mechanism that is at least in part independent of the killing function of *ced-9*, as *ced-9(lf); ced-3(lf)* engulfment-defective mutants were more defective in killing than were *ced-9(lf); ced-3(lf)* mutants (Table 3).

Engulfment-promoted killing also acts at least in part independently of the killing activity of *ced-8*, which controls the timing of cell death and, like engulfment, contributes to cell killing downstream of or parallel to *ced-9* protection²². Specifically, *ced-1(e1735)*

Table 3 Engulfment killing acts downstream of *ced-9* and parallel to *ced-8*

Genotype	Percentage of animals with undead V5R.paapp cells*	Number of extra cells (anterior pharynx)†
<i>ced-8(n1891)</i>	0	1.1 ± 0.17
<i>ced-1(e1735); ced-8(n1891)</i>	4	1.7 ± 0.22 (<i>n</i> = 40)
<i>ced-3(n2427)‡</i>	0	1.5 ± 0.14 (<i>n</i> = 47)
<i>ced-1(e1735); ced-3(n2427)‡</i>	48	5.9 ± 0.36
<i>ced-3(n2427); ced-8(n1891)‡</i>	0	5.7 ± 0.29 (<i>n</i> = 45)
<i>ced-1(e1735); ced-3(n2427); ced-8(n1891)‡§</i>	92	8.4 ± 0.31 (<i>n</i> = 45)
<i>ced-9(n2812); ced-3(n2427)</i>	0	6.3 ± 0.46
<i>ced-9(n2812); ced-3(n2427); ced-8(n1891)</i>	4	9.3 ± 0.40
<i>ced-1(e1735); ced-9(n2812); ced-3(n2427)</i>	80	9.8 ± 0.36

* The number of cells in the right postdeirid of L4 larvae was determined using Nomarski optics. In the wild type four cells are present, and one cell undergoes programmed cell death. Data are percentages. Sample size *n* = 25.

† The number of extra cells in the anterior pharynxes of L3 larvae was determined using Nomarski optics. In the wild-type, 16 cells undergo programmed cell death in this region. Data are means ± standard error of the means. Sample size *n* = 30 unless otherwise shown.

Strains also contained the following mutations: ‡ *unc-30(e191)*; § *sem-4(n1378)*.

enhanced the killing defect of *ced-3(n2427)* in the absence of *ced-8* (Table 3) ($P < 0.0001$, unpaired *t*-test). We suggest that engulfment, like *ced-8*, defines one of multiple downstream and partially redundant cell-killing activities.

The death of at least one cell type that normally dies during *C. elegans* development may be completely dependent on engulfment. The male-specific left-right homologues B.alapaav and B.arapaav can survive in engulfment-defective mutant animals³ and can also survive if their engulfing cell, P12.pa, is ablated using a laser microbeam²³. We confirmed that B.alapaav/B.arapaav dies in an engulfment-dependent manner and found that this death requires the function of the CED-3 caspase (data not shown). That engulfment can lead to a caspase-dependent death suggests that engulfing cells can trigger a signal transduction pathway that contributes to caspase activation and perhaps other aspects of programmed cell death. Consistent with the hypothesis that engulfing cells can signal changes within dying cells, the activities of two engulfment genes necessary for cell-corpse recognition, *ced-1* and *ced-7* (ref. 6), are required for the generation of free DNA 3'-hydroxyl ends within dying cells, a process that can occur without cell-corpse engulfment²⁴ and that is a hallmark of apoptotic cells²⁵.

We propose that engulfing cells promote the suicides of many and possibly all cells triggered to initiate programmed cell death. As described above, in the case of B.alapaav/B.arapaav an engulfing cell is necessary for death. As in *C. elegans*, in mammals dying cells are rapidly recognized by other cells and engulfed. The removal or

Table 2 Engulfment mutations result in the survival of cells that normally die in the ventral cord

Genotype	Number of corpses (range)*	Average number of extra VC-like cells†	Percentage of animals with surviving Pn.aap cells‡
<i>P_{lin-11gfp}</i>	0.0 ± 0.0	0.0 (<i>n</i> = 50)	0
<i>ced-3(n2452); P_{lin-11gfp}</i>	0.0 ± 0.0	5.0 (<i>n</i> = 50)	100
<i>ced-1(n1995); P_{lin-11gfp}</i>	6.3 ± 3.9	0.08 (<i>n</i> = 123)	7
<i>ced-1(n1951); P_{lin-11gfp}</i>	20.8 ± 4.1	0.21 (<i>n</i> = 100)	21
<i>ced-1(e1735); P_{lin-11gfp}</i>	24.8 ± 3.0	0.36 (<i>n</i> = 150)	31
<i>ced-2(n1994); P_{lin-11gfp}</i>	19.9 ± 4.2	0.35 (<i>n</i> = 100)	29
<i>ced-5(n2098); P_{lin-11gfp}</i>	24.2 ± 3.8	0.24 (<i>n</i> = 243)	21
<i>ced-5(n1812); P_{lin-11gfp}</i>	31.2 ± 3.1	0.44 (<i>n</i> = 100)	35
<i>ced-6(n1813); P_{lin-11gfp}</i>	20.2 ± 3.1	0.57 (<i>n</i> = 150)	45
<i>ced-7(n1996); P_{lin-11gfp}</i>	35.0 ± 5.7§	0.95 (<i>n</i> = 253)	67
<i>ced-7(n2001); P_{lin-11gfp}</i>	32.6 ± 3.6§	1.47 (<i>n</i> = 100)	79
<i>ced-7(n1892); P_{lin-11gfp}</i>	42.9 ± 3.8§	1.42 (<i>n</i> = 150)	80
<i>ced-10(n1993); P_{lin-11gfp}</i>	15.5 ± 4.0	0.24 (<i>n</i> = 150)	24
<i>ced-10(n3246); P_{lin-11gfp}</i>	29.6 ± 4.7	0.47 (<i>n</i> = 135)	31
<i>ced-12(n3261); P_{lin-11gfp}</i>	24.5 ± 3.7	0.31 (<i>n</i> = 100)	31
<i>ced-1(e1735); ced-2(n1994); P_{lin-11gfp}</i>	37.7 ± 4.7	0.48 (<i>n</i> = 127)	37

* Average numbers of persistent cell corpses were determined in the heads of at least 25 L1 larvae at the four-cell gonad stage using Nomarski microscopy. Error indicates standard deviation.

† Average numbers of fluorescent cells caused by expression from *P_{lin-11gfp}*, *nls96*, in P2, 9, 10, 11, and 12-derived regions were determined using a stereomicroscope equipped with an ultraviolet light source (Kramer Scientific).

‡ The percentages of animals that had at least one fluorescent cell in the P2, 9, 10, 11, and 12-derived regions were determined.

§ Cell corpses in the heads of *ced-7* mutant animals were counted in four-fold embryos, because corpse number in *ced-7* mutant larvae decreases sharply with time⁶.

|| *ced-1(e1735); ced-2(n1994)* double mutants had similar defects in cell-corpse engulfment in the ventral cords of *ced-1(e1735)* mutants (6.2 ± 0.2 (*n* = 30) and 6.1 ± 0.1 (*n* = 60) out of seven possible corpses in the posterior ventral cords, respectively).

inhibition of macrophages can disrupt the remodelling of tissues in the mouse eye or tadpole tail regression, processes that involve apoptosis^{26,27}. Furthermore, the elimination of macrophages in the anterior chamber of the rat eye results in the survival of vascular endothelial cells that normally undergo apoptosis²⁸. These findings suggest a potential role for macrophages in promoting the deaths of cells during tissue regression and suggest that engulfment may be involved in facilitating programmed cell death in some tissues in vertebrates. If so, the pharmacologic inhibition of engulfment in humans may be of therapeutic benefit in situations in which cells are poised between survival and death, such as in ischemic regions surrounding areas of infarction or in slow neurodegenerative diseases⁸. □

Methods

Strains and genetics

All strains were grown at 20 °C on NGM (nematode growth medium) agar with *Escherichia coli* OP50 bacteria. The wild-type strain was N2. The following alleles were used (the molecular nature of the alleles of cell-death genes is also listed): LGI: *sem-4(n1378)*; *ced-12(n3261)*, loss of function (Z. Zhou and H.R.H., unpublished results); *ced-1(e1735)*; *Q3750opal*⁶; *ced-1(n1951)*; *C501Y*⁶; *ced-1(n1995)* P124L⁶. LGIII: *ced-4(n3158)*; *S163F* (B. Hersh, S.C. and H.R.H., unpublished); *ced-4(n3195)*; *R63H* (S.C. and H.R.H., unpublished); *ced-4(n3312)*; *A394T* (B. Hersh, P.W.R. and H.R.H., unpublished); *ced-6(n1813)*; *deletion*¹⁴; *ced-7(n1892)*, loss of function¹³; *ced-7(n1996)*; *R50chre*¹³; *ced-7(n2001)*; *W1540ochre*¹³; *unc-69(e587)*; *ced-9(n2812)*; *Q46amber*²⁰. LGIV: *ced-2(e1752)*; *W153opal* G192E¹⁶; *ced-2(n1994)*; *R102opal*¹⁶; *ced-10(n1993)*; *V190G*¹⁶; *ced-10(n3246)*; *G60R*¹⁶; *him-8(e1489)*; *ced-5(n1812)*; *E28opal*¹⁵; *ced-5(n2098)*; *splice mutation*¹⁵; *unc-30(e191)*; *ced-3(n2427)*; *G474R*¹⁷; *ced-3(n2438)*; *G474R*¹⁷; *ced-3(n2443)*; *P400S*¹⁷; *ced-3(n2447)*; *S446L*¹⁷; *ced-3(n2452)*; *deletion*¹⁷. LGV: *him-5(e1490)*; *egl-1(n3331)*; *A59P* (S.C. and H.R.H., unpublished results); *nls96*. LGX: *ced-8(n1891)*; *W219amber*²²; *lin-15(n765ts)*; *nls106*. Extra cells and cell corpses were visualized using Nomarski optics. VC neurons expressing GFP were visualized using a dissecting light microscope equipped with an ultraviolet light source (Kramer Scientific). Cell lineage analysis was performed as previously described⁹.

Generation of the *P_{lin-11}gfp* reporter

To construct the *P_{lin-11}gfp* reporter we fused the *gfp* gene to an *Nsil* fragment of genomic DNA containing 5.2 kb of 5' sequence and the ATG of the *lin-11* gene¹⁸. Reporter DNA was injected at 50 ng μ L⁻¹ into *lin-15(n765)* mutants using the *lin-15(+)* plasmid pL15EK²⁹ as a rescuing co-injection marker. Non-Muv (that is, non-*Lin-15*) animals with bright expression of GFP in the VC neurons were irradiated with 4,000 rad from a cobalt source, and chromosomal integrants were identified as animals that stably transmitted the transgene to all progeny (that is, all progeny were non-Muv). The *nls96* and *nls106* integrants were backcrossed to N2 six times, and the integration sites were mapped to LGV and LGX, respectively.

Cell autonomous rescue

P_{col-10}ced-1::gfp, *P_{egl-1}ced-1::gfp*, and *P_{col-10}gfp* were previously described⁶. We injected these constructs at 75 ng μ L⁻¹ with *ttx-3::gfp* (ref. 30) at 50 ng μ L⁻¹ as the co-injection marker into *ced-1(e1735)*; *ced-3(n2427)*; *nls96* animals. Transgenic animals were identified by GFP fluorescence in the A1Y.

Received 19 February; accepted 8 May 2001.

- Ellis, H. M. & Horvitz, H. R. Genetic control of programmed cell death in the nematode *C. elegans*. *Cell* **44**, 817–829 (1986).
- Conradt, B. & Horvitz, H. R. The *C. elegans* protein EGL-1 is required for programmed cell death and interacts with the Bcl-2-like protein CED-9. *Cell* **93**, 519–529 (1998).
- Hedgecock, E. M., Sulston, J. E. & Thomson, J. N. Mutations affecting programmed cell deaths in the nematode *Caenorhabditis elegans*. *Science* **220**, 1277–1279 (1983).
- Ellis, R. E., Jacobson, D. M. & Horvitz, H. R. Genes required for the engulfment of cell corpses during programmed cell death in *Caenorhabditis elegans*. *Genetics* **129**, 79–94 (1991).
- Chung, S., Gumienny, T. L., Hengartner, M. O. & Driscoll, M. A common set of engulfment genes mediates removal of both apoptotic and necrotic cell corpses in *C. elegans*. *Nature Cell Biol.* **2**, 931–937 (2000).
- Zhou, Z., Hartwig, E. & Horvitz, H. R. CED-1 is a transmembrane receptor that mediates cell corpse engulfment in *C. elegans*. *Cell* **104**, 43–56 (2001).
- Yuan, J., Shaham, S., Ledoux, S., Ellis, H. M. & Horvitz, H. R. The *C. elegans* cell death gene *ced-3* encodes a protein similar to mammalian interleukin-1 beta-converting enzyme. *Cell* **75**, 641–652 (1993).
- Vaux, D. L. & Korsmeyer, S. J. Cell death in development. *Cell* **96**, 245–254 (1999).
- Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
- Robertson, A. M. G. & Thomson, J. N. Morphology of programmed cell death in the ventral nerve cord of *C. elegans* larvae. *J. Embryol. Exp. Morphol.* **67**, 89–100 (1982).
- Metzstein, M. M., Stanfield, G. M. & Horvitz, H. R. Genetics of programmed cell death in *C. elegans*: Past, present and future. *Trends Genet.* **14**, 410–416 (1998).
- Albert, M. L., Kim, J. I. & Birge, R. B. alpha V β 5 integrin recruits the CrkII-Dock180-rac1 complex for phagocytosis of apoptotic cells. *Nature Cell Biol.* **2**, 899–905 (2000).
- Wu, Y. C. & Horvitz, H. R. The *C. elegans* cell corpse engulfment gene *ced-7* encodes a protein similar

- to ABC transporters. *Cell* **93**, 951–960 (1998).
- Liu, Q. A. & Hengartner, M. O. Candidate adaptor protein CED-6 promotes the engulfment of apoptotic cells in *C. elegans*. *Cell* **93**, 961–972 (1998).
- Wu, Y. C. & Horvitz, H. R. *C. elegans* phagocytosis and cell-migration protein CED-5 is similar to human DOCK180. *Nature* **392**, 501–504 (1998).
- Reddien, P. W. & Horvitz, H. R. CED-2/CrkII and CED-10/Rac control phagocytosis and cell migration in *Caenorhabditis elegans*. *Nature Cell Biol.* **2**, 131–136 (2000).
- Shaham, S., Reddien, P. W., Davies, B. & Horvitz, H. R. Mutational analysis of the *Caenorhabditis elegans* cell-death gene *ced-3*. *Genetics* **153**, 1655–1671 (1999).
- Freyd, G., Kim, S. K. & Horvitz, H. R. Novel cysteine-rich motif and homeodomain in the product of the *Caenorhabditis elegans* cell lineage gene *lin-11*. *Nature* **344**, 876–879 (1990).
- Conradt, B. & Horvitz, H. R. The TRA-1A sex determination protein of *C. elegans* regulates sexually dimorphic cell deaths by repressing the *egl-1* cell death activator gene. *Cell* **98**, 317–327 (1999).
- Hengartner, M. O. & Horvitz, H. R. *C. elegans* cell survival gene *ced-9* encodes a functional homolog of the mammalian proto-oncogene bcl-2. *Cell* **76**, 665–676 (1994).
- Hengartner, M. O. & Horvitz, H. R. Activation of *C. elegans* cell death protein CED-9 by an amino-acid substitution in a domain conserved in Bcl-2. *Nature* **369**, 318–320 (1994).
- Stanfield, G. M. & Horvitz, H. R. The *ced-8* gene controls the timing of programmed cell deaths in *C. elegans*. *Mol. Cell* **5**, 423–433 (2000).
- Sulston, J. E., Albertson, D. G. & Thomson, J. N. The *Caenorhabditis elegans* male: postembryonic development of nongonadal structures. *Dev. Biol.* **78**, 542–576 (1980).
- Wu, Y. C., Stanfield, G. M. & Horvitz, H. R. NUC-1, a *Caenorhabditis elegans* DNase II homolog, functions in an intermediate step of DNA degradation during apoptosis. *Genes Dev.* **14**, 536–548 (2000).
- Wyllie, A. H. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation. *Nature* **284**, 555–556 (1980).
- Lang, R. A. & Bishop, J. M. Macrophages are required for cell death and tissue remodeling in the developing mouse eye. *Cell* **74**, 453–462 (1993).
- Little, G. H. & Flores, A. Inhibition of programmed cell death by catalase and phenylalanine methyl ester. *Comp. Biochem. Physiol. Physiol.* **105**, 79–83 (1993).
- Diez-Roux, G. & Lang, R. A. Macrophages induce apoptosis in normal cells in vivo. *Development* **124**, 3633–3638 (1997).
- Clark, S. G., Lu, X. & Horvitz, H. R. The *Caenorhabditis elegans* locus *lin-15*, a negative regulator of a tyrosine kinase signaling pathway, encodes two different proteins. *Genetics* **137**, 987–997 (1994).
- Hobert, O. *et al.* Regulation of interneuron function in the *C. elegans* thermoregulatory pathway by the *ttx-3* LIM homeobox gene. *Neuron* **19**, 345–357 (1997).

Acknowledgements

We thank B. Galvin, B. Hersh and Z. Zhou for comments. P.W.R. was supported by a National Science Foundation Fellowship. S.C. was supported by a Howard Hughes Medical Institute (HHMI) postdoctoral fellowship, a Merck/MIT collaborative fellowship, and an NIH training grant. H.R.H. is an Investigator of the HHMI.

Correspondence and requests for materials should be addressed to H.R.H. (e-mail: horvitz@mit.edu).

Engulfment genes cooperate with *ced-3* to promote cell death in *Caenorhabditis elegans*

Daniel J. Hoepfner^{*†‡}, Michael O. Hengartner^{*} & Ralf Schnabel[§]

^{*} Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

[†] Graduate Program in Genetics, Department of Molecular Genetics and Microbiology, SUNY at Stony Brook, Stony Brook, New York 11794, USA

[§] Institut fuer Genetik, TU Braunschweig, Spielmannstrasse 7, D-38106 Braunschweig, Germany

Genetic studies have identified over a dozen genes that function in programmed cell death (apoptosis) in the nematode *Caenorhabditis elegans*^{1–3}. Although the ultimate effects on cell survival or engulfment of mutations in each cell death gene have been extensively described, much less is known about how these mutations affect the kinetics of death and engulfment, or the interactions between these two processes. We have used four-dimensional-Nomarski time-lapse video microscopy to follow in detail how cell death genes regulate the extent and kinetics of

[‡] Present address: NINDS-LMB, Building 36, National Institutes of Health, Bethesda, Maryland 20892-4092, USA.

apoptotic cell death and removal in the early *C. elegans* embryo. Here we show that blocking engulfment enhances cell survival when cells are subjected to weak pro-apoptotic signals. Thus, genes that mediate corpse removal can also function to actively kill cells.

To determine the kinetics and morphological changes that occur during embryonic cell death, we analysed in detail the first 13 cells that die in the AB lineage (collectively referred to as the early

AB deaths; Supplementary information Table 1), using a four-dimensional-microscope time course analysis⁴. We found that a similar series of morphological changes occur during embryonic cell death as had been described during larval development^{5,6}. On the basis of studies using light microscopy, we divided embryonic cell death into five distinct steps: asymmetric cell division, ring stage, erythrocyte stage, lentil stage, and engulfment (see Supplementary Information Fig. 1).

Cell death during embryonic development is a highly stereotyped cell fate⁷. Complete reconstruction of the developmental lineage of three wild-type embryos showed that the same set of cells died in each case, with each cell dying at roughly the same point in development (R.S., unpublished observations; see also Supplementary Information and ref. 4). In contrast, we found that engulfment is not strictly defined by lineage: although some dying cells were repeatedly engulfed by the same eater in each of the three embryos that we recorded, other cells had variable fates (Supplementary Information Table 1). Thus, although we cannot exclude the possibility that lineage contributes to engulfment, cell identity is not sufficient to specify engulfing cell fate.

Seven genes (*ced-1*, -2, -5, -6, -7, -10 and -12) are known to participate in the removal of apoptotic cells in *C. elegans*³. However, none of these genes, taken individually, is essential for engulfment: only a fraction (usually less than one-third) of dying cells result in persistent cell corpses in any single mutant background⁸ (Fig. 1). To determine the basis of this low expressivity, we followed the early AB deaths in engulfment-defective mutants carrying candidate null mutations in either *ced-6* or *ced-7*. The *ced-6* gene encodes a phosphotyrosine-binding, domain-containing adaptor protein that mediates response to apoptotic cell recognition in engulfing cells⁹. *ced-7* encodes an ABC transporter similar to human ABCA1, and is required in both dying and engulfing cells¹⁰.

We found that about 20 and 40%, respectively, of the early AB deaths failed to be engulfed, and generated persistent cell corpses in *ced-6* and *ced-7* mutant embryos (Fig. 1). Persistence of cell corpses appeared to distribute randomly among all early AB deaths, both within and between the two mutant genotypes (see Supplementary Information Table 1). Thus, although we cannot exclude the

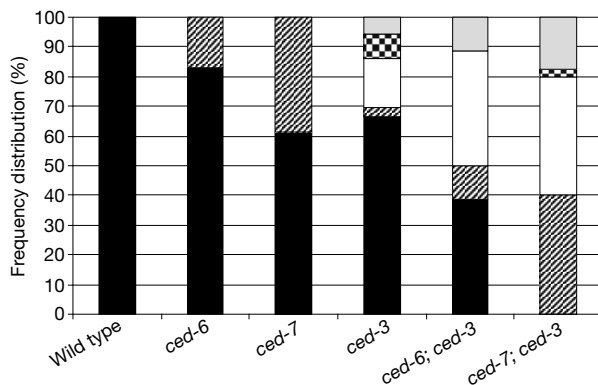


Figure 1 Fate distribution of early AB lineage deaths in wild-type and mutant backgrounds. The fates of the early AB lineage cell deaths were followed in three embryos. Black bars indicate normal cell deaths where all morphological stages were identified. White bars indicate cells that didn't display any morphological evidence of cell death: all such cells survived. Hatched bars indicate normal cell deaths that failed to be engulfed and formed persistent cell corpses. Chequered bars indicate cells that initiated morphological apoptosis but were engulfed without ever reaching the lentil stage. Shaded bars indicate reversion events where cells displayed early morphological hallmarks of death, but reverted to a normal morphology and survived until the end of the recording period. Genotypes and number of cells scored: wild type ($n = 35$); *ced-6*($n1813$) ($n = 35$); *ced-7*($n1892$) ($n = 31$); *ced-3*($op149$) ($n = 36$); *ced-6*($n1813$); *ced-3*($op149$) ($n = 36$); *ced-7*($n1892$); *ced-3*($op149$) ($n = 45$). A detailed description of the fate of each cell, as well as representative movies showing the various cell fates described here, are available as Supplementary Information.

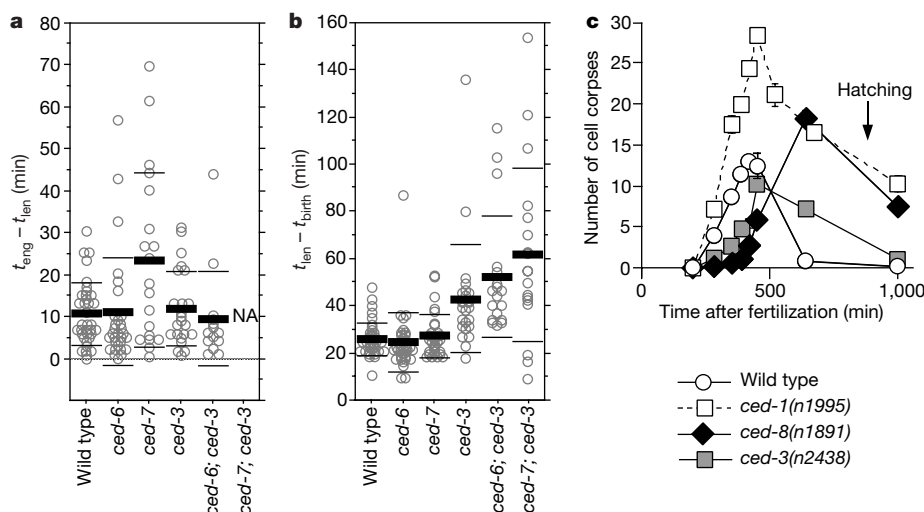


Figure 2 Mutations in *ced* genes affect the kinetics of killing and engulfment. **a**, Latent period ($t_{\text{eng}} - t_{\text{len}}$) in cells that both die and are engulfed. t_{eng} , time of engulfment; t_{len} , time at lentil stage, NA, not applicable. None of the cells in the *ced-7* *ced-3* double mutant lineages were engulfed, they all either survived or formed persistent cell corpses. Genotypes and number of cells scored: wild type ($n = 36$); *ced-6*($n1813$) ($n = 30$); *ced-7*($n1892$) ($n = 19$); *ced-3*($op149$) ($n = 23$); *ced-6*($n1892$); *ced-3*($op149$) ($n = 14$). **b**, Reduction in *ced-3* activity results in delayed cell death. The time elapsed between birth and full-blown apoptosis (lentil stage) was plotted for the early AB deaths. For **a** and

b, each observation is represented by a grey circle. Black bars indicate mean $\pm$ s.d. Genotypes and number of cells scored: wild type ($n = 36$); *ced-6*($n1813$) ($n = 35$); *ced-7*($n1892$) ($n = 31$); *ced-3*($op149$) ($n = 25$); *ced-6*($n1813$); *ced-3*($n2438$) ($n = 18$); *ced-7*($n1892$); *ced-3*($op149$) ($n = 18$). **c**, Reduction in *ced-3* function results in delayed embryonic cell death. Cell corpses were scored in embryos at various developmental stages. Data represent mean $\pm$ s.e.m. Error bars are within the symbols, except where shown.

possibility that some early AB deaths are more sensitive to the absence of *ced-6* and/or *ced-7* function than others, our limited data set is consistent with the suggestion that failure to be engulfed is essentially a stochastic process in the early embryo. Notably, *ced-7* not only affected the likelihood of a cell being engulfed, but also the efficiency of that process: even the cell corpses that were ultimately engulfed persisted almost twice as long in *ced-7* mutants than in the wild type (Fig. 2a). In contrast, we did not observe any delay in engulfment in *ced-6(n1813)* mutants, possibly reflecting a different mode of action of the two genes.

Loss-of-function mutations in the caspase *ced-3*, the Apaf-1 homologue *ced-4*, and the BH3 domain gene *egl-1*, as well as a gain-of-function mutation in the *bcl-2* homologue *ced-9*, prevent embryonic cell death in *C. elegans* (reviewed in ref. 1). As expected, we found that the strong mutation *ced-3(n717)* extensively suppressed death in the early AB lineage (data not shown). In contrast, the weaker mutation *ced-3(op149)* only partially suppressed the early AB deaths (Fig. 1). Mutations in *ced-3* not only affect the probability of survival, but also the kinetics of death of the cells that do die: onset of death (lentic stage) was delayed by an average of 75% in doomed *ced-3(op149)* cells when compared to wild type (Fig. 2b). These results suggest that initiation of death requires a certain threshold of caspase substrate cleavage. Weak *ced-3* mutants, which have lower caspase activity, might need more time to reach this threshold. A similar delay in the initiation of apoptosis has been described in *ced-8* mutants¹¹. To determine whether delayed cell death is a specific feature of the early AB deaths or a more general phenomenon, we scored the number of cell corpses visible at different times during embryonic development. We found that *ced-3(n2438)* mutants showed a delayed wave of embryonic cell deaths, similar to the one observed in *ced-8* mutants (Fig. 2c). These results are consistent with the hypothesis that cell death is generally delayed in weak *ced-3* mutants.

In addition to the two expected phenotypic classes (cell death followed by engulfment, or cell survival), we observed, at a low frequency, three other cell fates in *ced-3(op149)*. In one case, a cell died but failed to be engulfed and generated a persistent cell corpse (as in an engulfment-defective mutant), suggesting that caspase activity is required—directly or indirectly—for efficient activation of the pro-engulfment signal(s).

Furthermore, two cells proceeded through the early morphological stages of cell death (ring and erythrocyte stages), but then

reverted to the appearance of normal cells and survived (Fig. 1). These observations hinted at the possibility that, under reduced caspase activity, cells have the ability to reverse course and escape death, even after activation of the apoptotic programme. Because activation of CED-3 is required for the early morphological features of cell death, these observations suggest that, at least under our experimental conditions, cells can activate CED-3 and still escape death. How cells accomplish this feat is unknown; *C. elegans* may contain inhibitors of active caspases—analogs of the mammalian IAP proteins¹² or even CED-9 itself¹³—which could arrest a weak activation process.

One factor that might limit the ability of a cell to fully revert is the activation of the engulfment programme, which would result in an early or 'quasi' apoptotic cell being engulfed and eliminated. Indeed, in three cells, cell death was initiated but then proceeded to engulfment without going through the lentic stage that is characteristic of apoptotic cell corpses. These observations suggest that engulfment can occur independently of full-blown morphological apoptosis. We suggest that active CED-3 initiates a number of downstream subprogrammes (for example, DNA degradation, cell shrinkage, emission of 'eat me' signals), which are responsible for the orderly dismantling and removal of apoptotic cells. At low caspase levels, coordinated execution of these subprogrammes might be impaired. Uncoupling of downstream events by altering the kinetics of death has also been observed in *ced-8* mutants¹¹, as well as in mammalian cells treated with caspase inhibitors.

Our observations suggest that in a weak *ced-3* background, cells at early stages (ring and erythrocyte stages) of cell death have three options: progression to full-blown apoptotic cell corpses, direct engulfment without morphological progression to full corpse, or reversion to normality and survival. We proposed that the relative frequency of the latter two fates might be influenced by the efficiency with which the early corpse is recognized and engulfed. To determine whether preventing engulfment might enhance reversion from 'early death' and hence cell survival, we followed the fate of the early AB cells in *ced-6(n1813); ced-3(op149)* and *ced-7(n1892); ced-3(op149)* double mutant embryos.

Indeed, we found that the absence of *ced-6* or *ced-7* function significantly increased the frequency of such reversion events (Fig. 1). Furthermore, the fraction of cells that never initiated any overt signs of apoptosis was also greatly increased (Fig. 1). The counterintuitive conclusion from our results is that interfering with

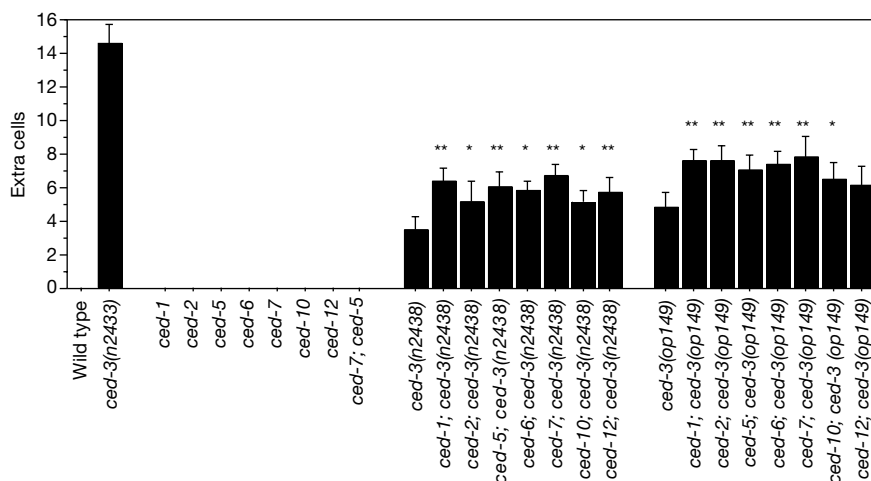


Figure 3 Engulfment promotes cell death in weak *ced-3* mutant backgrounds. Number of extra cells in the anterior pharynx of control animals and engulfment-defective mutants alone or in combination with weak alleles of *ced-3*. In comparing *ced-3* single mutants with double-mutant combinations, asterisks indicate $P < 0.02$ and double asterisks indicate $P < 0.001$ (paired t -test, Statview version 4.5 software). Error bars reflect 95% confidence limits ($n = 20$ animals for each genotype). All animals were scored as L3 or L4

larvae. The engulfment genes enhance mutations in *ced-3* far more strongly than in its upstream regulators (See Supplementary Information Fig. 2). This differential behaviour might simply reflect the closer genetic proximity of CED-3 to the engulfment pathway. Alternatively, it might reflect different kinetics of cell-death activation in these mutant backgrounds. Alleles used: *ced-1(e1735)*, *ced-2(e1752)*, *ced-5(n1812)*, *ced-6(n1813)*, *ced-7(n1792)*, *ced-10(n1993)* and *ced-12(oz167)*.

the removal of apoptotic (presumably dead) cells not only allows 'early dead' cells to revert and escape death, but also results in many cells not dying (becoming apoptotic) in the first place.

Do the reversions observed in the absence of engulfment simply correspond to a suspended sentence, with all such cells eventually dying but at a later time point, or do they result in permanent survival of the affected cells? Because four-dimensional microscopic analysis cannot readily be continued past the first half of embryonic development, we resorted to measuring the long-term survival of a group of 16 pharyngeal cells that normally die by apoptosis during embryonic development^{14,15}. Mutations that only transiently delay cell death do not score strongly in this assay¹¹. In contrast, we found that about 50% more cells survived in *ced-6(n1813); ced-3(op149)* and *ced-7(n1892); ced-3(op149)* double mutants than in the *ced-3(op149)* single mutant. This effect is not specific to the *op149* allele, as we obtained similar results with another weak *ced-3* mutation, *n2438* (Fig. 3). Thus, elimination of engulfment gene function results in long-term cell survival.

To determine whether this enhanced survival is a general feature of mutations in the engulfment pathway, we tested the effect of the other known engulfment genes on cell survival. As expected, engulfment-defective single mutants did not show any increased cell survival in the anterior pharynx on their own (Fig. 3). In contrast, double-mutant combinations of any engulfment-defective mutation with weak alleles of *ced-3* significantly enhanced cell survival (Fig. 3). These results suggest that the increased survival stems from interfering with the engulfment pathway, rather than being a specific property of *ced-6* and *ced-7*.

How can the engulfment machinery contribute to cell killing? Our data are consistent with at least two models (Fig. 4). The 'backup-plan model' suggests that low levels of CED-3 caspase activity might on occasion be sufficient to activate the eat-me signal on the surface of the cell, but not enough to kill the cell. However, even if the cell does not autonomously kill itself, exposure of the eat-me signal ensures that it will be recognized and engulfed, and therefore properly removed. The alternative 'positive-feedback

model' proposes that doomed cells indicate their desire to die to neighbouring cells, either through cell-surface changes or secretion of a signalling molecule. Recognition of this signal results in the neighbouring cells sending back pro-apoptotic signals, encouraging the doomed cell to 'go for it', thereby ensuring that the cell completes the process. Given our data, completion of this positive feedback loop would somehow require an intact engulfment pathway, possibly for transduction of the signal in the neighbouring cells.

Our experimental data do not currently allow us to distinguish between these two models. Our observation of at least a few cases of engulfment before normal corpse formation supports the existence of the backup-plan model. In mammals, it has been shown that macrophages could bind in a caspase-dependent manner to neutrophils that are morphologically normal¹⁶, also suggesting that the eat-me signals can be activated independently of the rest of the apoptotic programme. On the other hand, the fact that we observed a significant delay in the onset of morphological cell death in our double mutants (Fig. 2b) strongly supports the positive-feedback model. Experiments in mammals have also shown that engulfing cells can promote the death of their potential target. Macrophage-induced apoptosis has been observed under various conditions^{17,18} and in several cases, the molecular basis of death signalling has been at least partially elucidated¹⁹.

The pro-apoptotic function of the engulfment machinery is only noticeable under conditions of limiting caspase activity (weak *ced-3* mutants). One possible explanation is that this function is preferentially activated after weak pro-apoptotic signals. Alternatively, the pro-apoptotic function of the engulfment pathway might also be present under normal conditions but redundant with the killing machinery, perhaps because it is a slower process. According to this hypothesis, *ced-3* promotes cell death with such a high efficiency in wild-type cells that the engulfment pathway is not given any time to act. However, when *ced-3* function is reduced, the kinetics of killing are slowed down (Fig. 2b), thereby allowing the engulfment genes the opportunity to make a detectable contribution.

The coordination of cell death with engulfment prevents dead

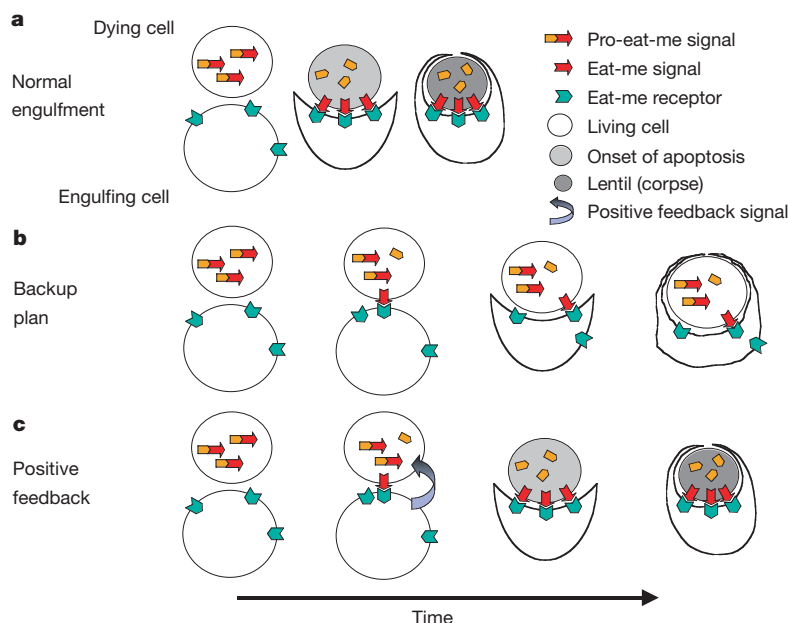


Figure 4 Possible modes of interaction between killing and engulfment. **a**, Normal engulfment involves the presentation of eat-me signals on the surface of a dying cell. The signal is presented as ligands to which receptors on the surface of the engulfing cell bind. Engagement of the receptor leads to cytoskeletal rearrangements within the engulfing cell and extension of pseudopods around the dying cell. **b**, Backup plan. This model postulates that apoptosis consists of several, independent subprogrammes. Activation of the engulfment subprogramme is sufficient to remove the cell, even if the cell-autonomous

subprogrammes were not properly initiated. The longer time course reflects the slower kinetics of this type of death. **c**, Positive feedback. In this model, activation of caspases in the doomed cell is sensed by the neighbouring cell, using the same set of genes that are used for engulfment. The neighbouring cell responds by emitting feedback signals, which stimulate the doomed cell to complete the apoptotic programme. We assume that under normal conditions, this positive-feedback mechanism is too slow to influence the death process.

cells from undergoing secondary necrosis and releasing potentially harmful cytosolic contents onto healthy neighbouring cells. Here we show that the engulfment programme also promotes the elimination of cells that are marginally viable. Because of the conserved function of much of the cell-death pathway between *C. elegans* and vertebrates, we predict that the ability of the phagocytic machinery to contribute to killing may be a general phenomenon in metazoans. The ability of the engulfment machinery to promote death in mammals might have important clinical implications. Such a programme might contribute to the elimination of cells under pathological conditions, in which cells are exposed to sublethal pro-apoptotic stimuli. Such conditions can often be found in chronic neurodegenerative diseases, after a stroke, as well as in developing tumours. Modulating the activity of the engulfing machinery in neighbouring cells might therefore offer a new target for therapies to treat these diseases. □

Methods

Mutations and strains

Animals were cultured on agar worm plates spotted with *Escherichia coli* strain OP50 at 20 °C, unless otherwise stated, essentially as described²⁰. We used the *C. elegans* N2 Bristol ecotype as the wild-type strain. The mutations used in this study are: linkage group (LG) I: *ced-12(oz167)*, *ced-1(e1735, n1995)*; LG III: *ced-4(n2273cs)*, *ced-6(n1813)*, *ced-7(n1892)*, *ced-9(n1950sd)*; LG IV: *ced-2(e1752)*, *ced-10(n1993)*, *ced-5(n1812)*, *ced-3(n2433, n2438, op149)*; LG V: *egl-1(n1084sd)*; LG X: *ced-8(n1891)*. The mutation *ced-3(op149)* was isolated as a recessive suppressor of the egg-laying defect present in *egl-1(n1084dm)* (D.J.H. and M.O.H., unpublished observations), and results in a V(GTT) to I(ATT) substitution at residue 426. The mutations *ced-12(oz167)*, *ced-3(n2433)* and *ced-3(n2438)* are described in refs 21, 15 and 22, respectively. All other mutations are described in ref. 23.

Video recording

Microscopic analysis of cells was carried out using standard live-animal mounting techniques on a Zeiss Axioplan microscope fitted with Nomarski optics. Embryo lineages were determined as described⁴.

Cell counts

Pharyngeal cell survival and embryonic cell corpses were scored as described¹⁴. We performed all statistical analyses using the Statview program (version 4.5; Abacus Concepts).

Received 14 March; accepted 14 June 2001.

1. Metzstein, M. M., Stanfield, G. M. & Horvitz, H. R. Genetics of programmed cell death in *C. elegans*: past, present and future. *Trends Genet.* **14**, 410–416 (1998).
2. Hengartner, M. O. in *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 383–415 (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1997).
3. Gumienny, T. L. & Hengartner, M. How the worm removes corpses: the nematode *C. elegans* as a model system to study engulfment. *Cell Death Diff.* (in the press).

4. Schnabel, R., Hutter, H., Moerman, D. & Schnabel, H. Assessing normal embryogenesis in *Caenorhabditis elegans* using a 4D microscope: variability of development and regional specification. *Dev. Biol.* **184**, 234–265 (1997).
5. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
6. Robertson, A. & Thomson, N. Morphology of programmed cell death in the ventral nerve cord of *Caenorhabditis elegans* larvae. *J. Embryol. Exp. Morph.* **67**, 89–100 (1982).
7. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
8. Ellis, R. E., Jacobson, D. M. & Horvitz, H. R. Genes required for the engulfment of cell corpses during programmed cell death in *Caenorhabditis elegans*. *Genetics* **129**, 79–94 (1991).
9. Liu, Q. A. & Hengartner, M. O. Candidate adaptor protein CED-6 promotes the engulfment of apoptotic cells in *C. elegans*. *Cell* **93**, 961–972 (1998).
10. Wu, Y.-C. & Horvitz, H. R. The *C. elegans* cell-corpse engulfment gene *ced-7* encodes a protein similar to ABC transporters. *Cell* **93**, 951–960 (1998).
11. Stanfield, G. M. & Horvitz, H. R. The *ced-8* gene controls the timing of programmed cell deaths in *C. elegans*. *Mol. Cell* **5**, 423–433 (2000).
12. Miller, L. K. An exegesis of IAPs: salvation and surprises from BIR motifs. *Trends Cell Biol.* **9**, 323–328 (1999).
13. Xue, D. & Horvitz, H. R. Inhibition of the *Caenorhabditis elegans* cell-death protease CED-3 by a CED-3 cleavage site in baculovirus p35 protein. *Nature* **377**, 248–251 (1995).
14. Hengartner, M. O., Ellis, R. E. & Horvitz, H. R. *C. elegans* gene *ced-9* protects cells from programmed cell death. *Nature* **356**, 494–499 (1992).
15. Shaham, S., Reddien, P. W., Davies, B. & Horvitz, H. R. Mutational analysis of the *Caenorhabditis elegans* cell-death gene *ced-3*. *Genetics* **153**, 1655–1671 (1999).
16. Knepper-Nicolai, B., Savill, J. & Brown, S. B. Constitutive apoptosis in human neutrophils requires synergy between calpains and the proteasome downstream of caspases. *J. Biol. Chem.* **273**, 30530–30536 (1998).
17. Lang, R., Lustig, M., Francois, F., Sellinger, M. & Plesken, H. Apoptosis during macrophage-dependent ocular tissue remodelling. *Development* **120**, 3395–3403 (1994).
18. Diez-Roux, G. & Lang, R. A. Macrophages induce apoptosis in normal cells *in vivo*. *Development* **124**, 3633–3638 (1997).
19. Duffield, J. S. *et al.* Activated macrophages direct apoptosis and suppress mitosis of mesangial cells. *J. Immunol.* **164**, 2110–2119 (2000).
20. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
21. Chung, S., Gumienny, T. L., Hengartner, M. O. & Driscoll, M. A common set of genes mediate removal of both apoptotic and necrotic cell corpses in *C. elegans*. *Nature Cell Biol.* **2**, 931–937 (2000).
22. Hengartner, M. O. & Horvitz, H. R. Activation of *C. elegans* cell death protein CED-9 by an amino-acid substitution in a domain conserved in Bcl-2. *Nature* **369**, 318–320 (1994).
23. Hodgkin, J. in *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 881–1047 (Cold Spring Harbor Laboratory, Plainview, 1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy form the London editorial office of *Nature*.

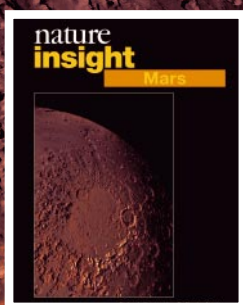
Acknowledgements

We thank P. Reddien and R. Horvitz for sharing results before publication, and G. Chimini, V. Fadok, N. Franc, R. Lang, J. Savill, P. Williamson, and members of the Hengartner laboratory for useful comments. This work was supported by an NIH grant and a research grant from Devgen NV to M.O.H.

Correspondence and requests for materials should be addressed to M.O.H. (e-mail: michael.hengartner@molbio.unizh.ch).

nature insight

Mars



Cover illustration

The Argire Basin on Mars, adapted from an image combining topography from the Mars Orbiter Laser Altimeter and a Viking colour image mosaic. [Image courtesy of MOLA Science Team and G. Shirah, NASA/GSFC Scientific Visualization Studio.]

This summer, Mars appears larger and brighter to us than it has for over a decade. We are reminded that each time we take a closer look at the red planet, the new observations seem only to increase our sense of wonder at how many mysteries remain unresolved regarding this most Earth-like of the other planets in our Solar System. At the heart of our fascination with Mars, at least over the past century, has been the possibility that liquid water exists at or near the martian surface and therefore that life might have taken hold elsewhere in our Solar System — or even exist there today.

And so, on the heels of some dramatic successes (and failures) in our attempts to deploy probes to study the martian surface, the time is appropriate to take stock of our understanding of present-day Mars as well as the geological and environmental evolution that has taken place since its formation, some four billion years ago. Most of our information on the interior of Mars comes to us from remote sensing of the martian surface and subsequent inferences on how the surface has evolved over geological timescales in response to both geodynamic processes in the martian interior and erosional processes at its surface. At the heart of this *Nature Insight*, a series of interrelated Review Articles describes this evolution, from the differentiation and solidification of the martian core to the dynamics of its atmosphere.

But first, on page 209, Kevin Zahnle provides a historical introduction of how both the public and scientific view of Mars has evolved since the latter part of the nineteenth century and how these ideas continue to propel our interest in Mars to this day. The core of Mars and scenarios for how the martian dynamo evolved are described by David Stevenson on page 214 — with consequences as far reaching as the preservation of the early atmosphere of Mars. Maria Zuber, on page 220, then reviews insights that have been gained into the evolution of the mantle and core of Mars, much of which have arisen from data obtained by the spectacular Mars Global Surveyor mission. The martian surface is thought to have evolved in response to surface processes, primarily through the action of water, ice and wind, as Victor Baker describes on page 228. Bruce Jakosky and Roger Phillips next discuss (on page 237) what such observations can tell us about the martian hydrological cycle and what constraints can be put on the possibility of liquid water having existed on Mars for significant periods of time. Atmospheric dynamics and weather on Mars differ from that of the Earth, as Conway Leovy explains on page 245. And, in a final commentary on page 250, Michael Carr and James Garvin describe the plethora of future missions planned for Mars.

At the next opposition of Mars in 2003, while a new set of landers and rovers will be hurling towards Mars, we will be able to gaze upon the brightest and biggest Mars in the night sky than at any time in the past 1,000 years as we ponder what new surprises await us from the martian surface.

We are pleased to acknowledge the financial support of The Planetary Society in producing this *Insight*. As always, *Nature* carries the sole responsibility for all editorial content and peer-review.

John VanDecar Senior Editor

Karl Ziemelis Physical Sciences Editor



Decline and fall of the martian empire

Kevin Zahnle

NASA Ames Research Center, Moffett Field, California 94035-1000, USA

“Are they worlds, or are they mere masses of matter? Are physical forces alone at work there or has evolution begotten something more complex, something not unakin to what we know on Earth as life? It is in this that lies the peculiar interest of Mars.”

Percival Lowell (in ref. 1, p. 3)

Perhaps it was in a Japanese garden that Mr Percival Lowell, the already well-known American orientalist and travel writer, first understood with the peculiar force of revelation just why the discovery of canals on Mars was the great event of his time. Perhaps the news had come in the form of wadded newsprint cushioning gifts shipped from home; stories about base ball and canals on Mars wrapping a fruitcake, maybe. One sees Percival, 38 years old and losing his fascination with the East, stepping from stone to stone across a dry sand sea raked to evoke waves on open water and waves lapping the shores of bare rock islands. One sees him distractedly tracing channels in the sand with a stick, drawing geometric patterns. One imagines him forming a plan.

Lowell's peculiar revelation actually took place after he had returned home to Boston in late 1893 to arrange for the publication of *Occult Japan*, when he received as a Christmas gift Camille Flammarion's *La Planète Mars et ses Conditions d'Habitabilité* (ref. 2; see also ref. 3, p. 104). In this beautiful and heavily illustrated volume, Flammarion compiled essentially every credible telescope observation of Mars that had been made. He describes a Mars of dry plains and shallow seas, an obviously habitable world, but more

important, he also freely speculated of a Mars of canals built by a higher civilization than any ever known on Earth (ref. 2, p. 586). (Oddly, Flammarion himself was never able to see the canals.)

Lowell's latent interest in astronomy had been growing as his interest in Japan waned. He read Flammarion's book “with lightning speed” and scrawled upon it the imperative “Hurry!”³ He would need his own observatory, under clear, clean, still air, and quickly. The opposition of Mars in October 1894 would be the last good one for more than a decade.

The canals

The canali of Mars were discovered in 1877 by Giovanni Schiaparelli using a modest 22-cm refractor in Milan. Schiaparelli had first won fame by showing that the annual Perseid meteor shower was due to comet Swift-Tuttle³. This fame earned him his observatory, where he visually measured double stars to high accuracy. Thus his surprising 1877 map of Mars was taken seriously. The wonderfully evocative martian nomenclature (consider, for example, Tharsis, Chryse, Elysium, Amazonis, Trivium Charontis and Syrtis Major) that still survives derives from this map. The canali were at first depicted mostly as broad channels,

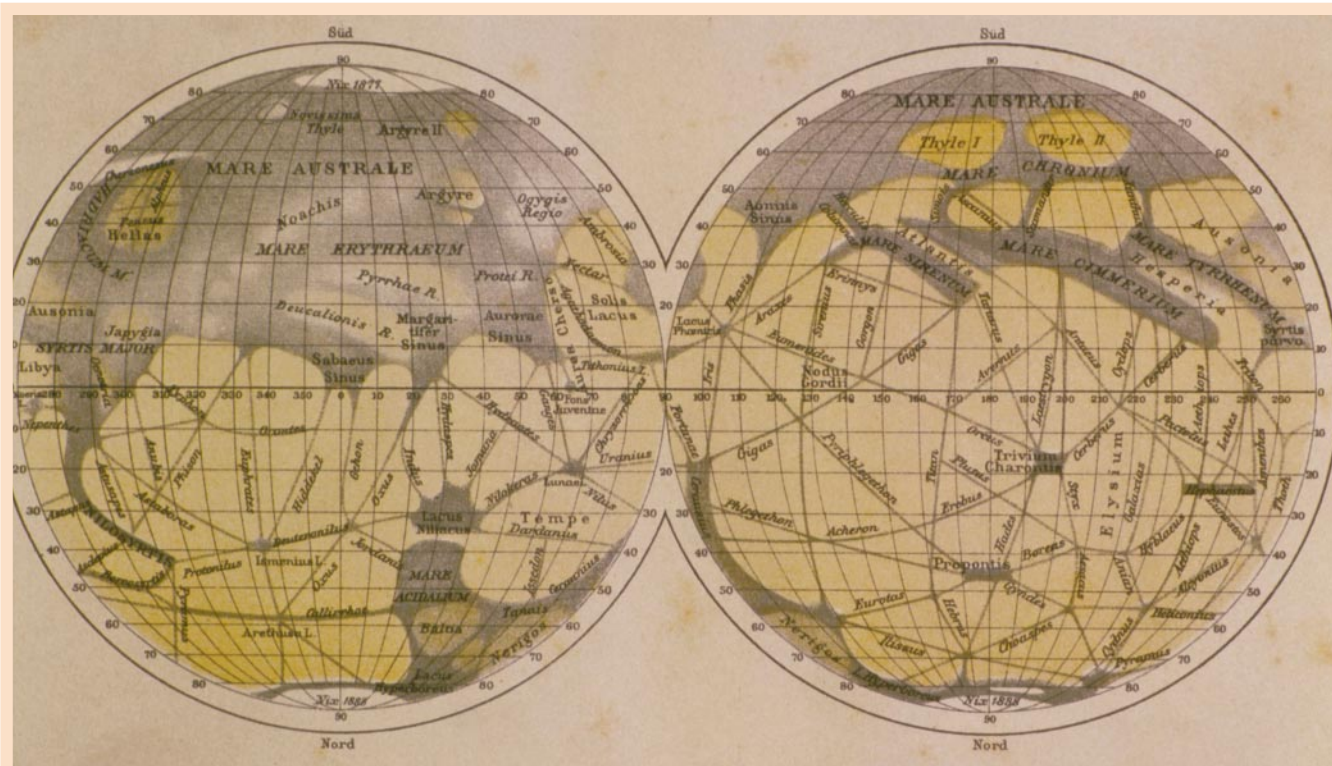


Figure 1 Mapping the surface of Mars. The Italian astronomer Giovanni Schiaparelli made this drawing in the early 1880s. He called the straight surface features canals, and also noticed that the patterns on the surface changed with the martian seasons, which he attributed to seasonal changes in vegetation.

more akin to the Malagasy Strait or the Red Sea than to the thin blue-green lines they were to become. (Schiaparelli's and other early areographies reflect the theory that the dark areas on Mars were seas; W. H. Pickering and Lowell showed the seas to be dry land.) At succeeding and progressively less favourable martian oppositions, Schiaparelli's canali hardened into a geometric network of razor-sharp, ruler-straight canals (Fig. 1). Many of the canali also began to appear in pairs, a controversial phenomenon that Schiaparelli referred to as "gemination".

At first, belief in the reality of the canals spread, as ever more observers seemed to glimpse some of what Schiaparelli had seen. In the late 1880s, apparent confirmations came from the world's largest refractors: Perrotin and Thollon in Nice, Schaerberle (and to a lesser extent Holden and Keeler) at Lick Observatory in California. Popular speculation centred on life and who might be living there, and what they might think of us. Signal lights were reported. An 1890 observation of projections towering above the martian limb (later shown to be dust clouds; ref. 4, p. 104) was also interpreted by some in the popular press as evidence of martians signalling Earth. A French widow, Clara Goguet Guzman, established a FF100,000 'Guzman prize' for the person or nation that first succeeded in establishing dialogue with another planet or star (ref. 3, p. 90), although it is said that she "excluded Mars because it would be too easy to establish contact" (ref. 5, p. 74). Later, in an episode preminiscent of the 'Face on Mars', one of Keeler's 1890 drawings of Mars was re-interpreted by the popular press as depicting the Hebrew letters for the name of God — "True, the magnitude of the work of cutting the canals into the shape of the name of God is at first thought appalling..." is a quote from the 2 June 1895 edition of the *San Francisco Chronicle* (ref. 3, p. 88).

A widespread story, perhaps spread by the anti-canalistas as some sort of apocryphal folk justice, is that Schiaparelli went blind. In fact, he lived long and prospered, with the favour of the King. Schiaparelli continued to observe Mars at every opposition until his death in 1910. However, he did not publish his observations of Mars made after 1890 because he regarded his vision as less keen than it had been formerly. More interesting were his coy delphic hints regarding

the artificiality, and towards the end of his life, even the very reality, of the canals. His contemporaries were never able to pin him down on the matter.

Lowell

Enter Percival Lowell. Lowell hired aides through his Harvard connections, including the confirmed canalista W. H. Pickering, and sent A. E. Douglass (who later founded the Steward Observatories at Tucson) to scout about Arizona and Mexico for good air at high elevation⁶. The setting of the Arizona city of Flagstaff on a high forested plateau fit Lowell's preconceptions of where the air would be best, and although in later years he would have cause to doubt his choice, time was of the essence.

The nature of Lowell's quest is revealed by what he said and wrote of the canals in the months before he first trained his new telescopes on Mars. Writing in the *Boston Commonwealth* on 26 May 1894, Lowell said: "The most self-evident explanation from the markings themselves is probably the true one; namely, that in them we are looking upon the result of the work of some sort of intelligent beings" (quoted by ref. 6, pp. 58–59).

His observatory was built and outfitted with borrowed 12- and 18-inch refractors in time for the martian season. Lowell and colleagues had to train themselves to see the canals (by all accounts they are hard to see, and as there were typically two or three observers who each tried to confirm what the others saw, there were other psychological factors at work), but once they learned how to do it they soon surpassed Schiaparelli's total.

Lowell's brilliant public lectures, his magazine articles, his tireless efforts at popularization that culminated in several books (*Mars*¹, published in 1895; *Mars and its Canals*⁴, published in 1906; and *Mars as an Abode of Life*⁷, published in 1908; see Box 1), made him and his Mars famous. In my opinion his argument can be reduced to the following essential points. (1) The polar caps are water ice. Here he chooses "between the rival candidates of common sense and uncommon subtlety, water and frozen carbonic acid gas" (ref. 4, p. 39). His key argument apart from "common sense" is the seasonal presence of

Box 1

Lowell's cosmogony

Lowell wrote three books concerned primarily with Mars: *Mars*¹ (1896); *Mars and Its Canals*⁴ (1906); and *Mars as an Abode of Life*⁷ (1908). The first bears the lightest touch, and is often marked by good humour. He both opens and closes with the classic arguments of SETI (the Search for Extraterrestrial Intelligence), but in the pages between he presents his case for Mars. He develops the case more fully in *Mars and Its Canals*. The book opens with an inspiring essay "On Exploration", in which he introduces the metaphor Carl Sagan later used for *Cosmos*. He divides the rest of the book into four parts: Natural Features, Non-Natural Features, The Canals in Action and Explanation. In this book Lowell shows signs of desperation. Although he is winning in the marketplace of public opinion (for example, *The Wall Street Journal* ranked "...the proof by astronomical observations...that conscious, intelligent human life exists upon the planet Mars" as one of the chief events of 1907), he is losing ground among scientists. One imagines that doubt gnawed at his sheets, but he betrays no uncertainty apart from less good humour. Large parts of the book are given over to dry observations of the canals, too painful now to read.

In *Mars as an Abode of Life*, Lowell invents "planetology" and expands on his cosmogony. He begins with nods towards the then fashionable ideas that the Universe consisted mostly of dark matter and that the Sun and its Solar System were born out of a close encounter between two dark stars. But his thoughts were mostly informed by a common sense (or folk science) understanding of surface-to-volume ratio. Worlds were formed hot. Large worlds cooled slowly, and were still evolutionarily young in 1894, "while in the moon we gaze upon the last sad age of decrepitude, a world almost sans air, sans sea, sans life, sans everything" (Lowell writing in the *Boston Commonwealth*, 26 May 1894; quoted in ref. 6, pp. 58–59). He places the origin of life in a Hadean realm of geothermal heat hidden from the Sun. This view of planetology remains in play.

In Lowell's planetology, worlds dry out as they age. One reason is that air escapes to space (discussed in detail in *Mars*, pp. 53–57), with the gases of low atomic mass, like water, escaping first. Another reason presumes cooling. "As the [internal] heat dissipates, the body begins to solidify, starting with the crust. For cosmic purposes it undoubtedly still remains plastic, but cracks of relatively small size are both formed and persist. Into these the surface water seeps. With continued refrigeration the crust thickens, more cracks are opened, and more water given lodgement within, to the impoverishment of the seas" (ref. 7, p. 146). Eventually the water that does not escape goes into hiding, leaving the surface a pitiless desert. If such a world still be inhabited, its inhabitants need husband their water. With some updating of his language (read cracking by impact cratering, let planetary cooling be manifest as little or no recycling, and let those inhabitants be our descendants) we obtain a contemporary picture of Mars.

a blue polar collar of meltwater that borders the retreating polar caps. At plausible pressures, dry ice does not liquefy. (2) A "wave of darkening" passes from pole to equator (and beyond) as the polar cap retreats. This is explained as growth of vegetation along canals and in oases as meltwater becomes available to more distant fields. (3) The "canaliform [sic] features" form a network too regular to be natural. Therefore they were constructed by an intelligence not unlike ours, for the purpose of husbanding and distributing scarce water. The aqueducts themselves would be unseen; rather the features represented swaths of irrigated land.

At first, Lowell's canals were received as credible, despite doubts regarding the theory by which he explained them, and despite the difficulty astronomers at Lick Observatory (a better site with a bigger telescope) had in confirming them. His location was good and his



Figure 2 Romance of the Heavens. Included in a cigarette-card set published in 1928 is an image depicting a distinctly Lowellian view of the martian landscape.

instruments were good. But confidence in Lowell fell sharply following his report in 1896 of a spoke-like pattern of linear markings on the surface of Venus⁶. By the opposition of 1907, despite a spectacular publicity campaign by Lowell and his staff featuring photographs of the greatest canals, many of the early canalistas had recanted, including in particular Antoniadi at Meudon, Cerulli at Collurania, and Douglass, his former assistant. Part of the problem was that the canals truly are not there; thus when conditions were especially good for observing them they were least visible. Antoniadi, who had begun as a protégé of Flammarion but with whom he later broke, reported seeing bewildering amounts of detail with Meudon's 33-inch refractor in 1909, but nothing at all that looked artificial. The rest of the problem was that Lowell's flamboyance, combativeness and inflexibility made him look ever more like a crank (Box 2).

What in fact were Schiaparelli, Lowell and others seeing? In *Mars*¹, pp. 145–147, Lowell lists 183 canals on Mars and the number of times each was seen in Flagstaff in 1894–5. The canal most often seen was Agathodaemon. When one compares Lowell's map to modern maps, one finds that Agathodaemon is Valles Marineris; that is, the most often seen canal on Mars coincides with the most prominent real channel on Mars. The second most often seen canal was Daemon, coincident with the chaos at the origin of Valles Marineris. A few other canals also prove real: for example, the double (geminated) canal Gigas corresponds to parallel rifts aligned with the great Tharsis volcanoes. But most of the canals are pure fancy. Lowell once said that "not twenty people in the world have seen them, but the fifteen who have seen them, have seen them." What we seem to have had here was a *folie à quinzé*.

Lowell's legacy

Pluto bears his name, but Mars is Percival Lowell's planet. By dint of his resources, talent and energy, and most of all by his choosing to "take the popular side of the most popular scientific question about" (W. W. Campbell, quoted by ref. 3 from Campbell's review of *Mars* in 1896), Lowell defined Mars in the popular imagination for nearly a century. His vision of a habitable and vegetated planet persisted into the space age, while his extreme position on the canals lent cover to those who would dream of a verdant Mars yet wished to think their hopes moderate.

Lowell's legacy reaches us today through two distributaries. One is through the mainstream of science, where the extent of his influence is emphasized by Horowitz⁸. Mid-century observers continued to address the thickness and composition of the martian atmosphere, the nature of the polar caps, the nature of the wave of darkening, and the nature of the dark stuff (that is, is it vegetation?). Because they were working at the edge of their resolution, most of the results they obtained were at the resolution limit. Unfortunately most of the results they obtained were wrong, because most of what they were looking for had signals well below their resolution limits. A brief overview of de Vaucouleurs' *The Physics of the Planet Mars*⁹ or the 1961 report of the Space Science Board¹⁰ to NASA shows how badly wrong they were: Mars possessed water-ice polar caps, 85-mbar N₂

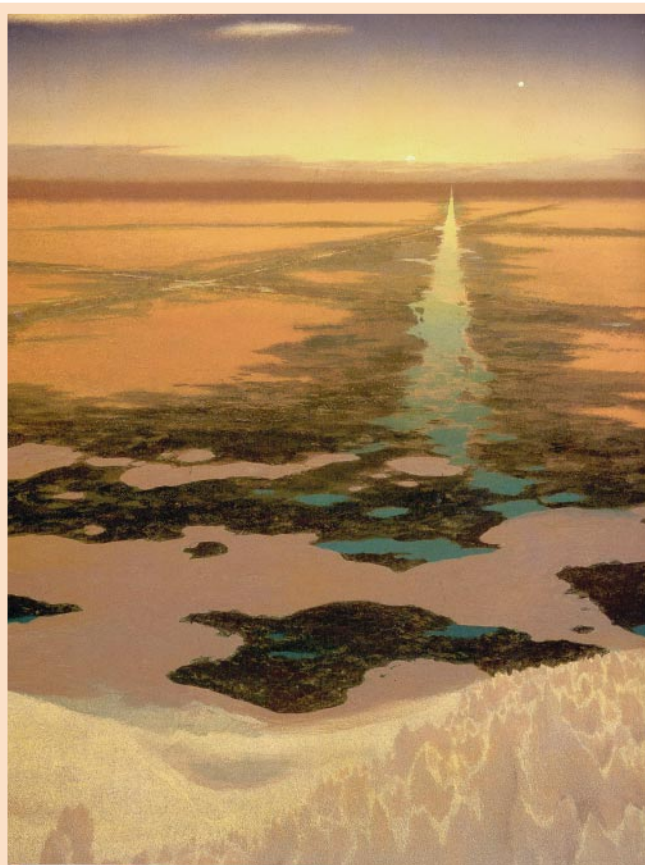


Figure 3 Lowell's legacy flows through popular culture, with martian canals and their builders seizing the public imagination.

atmosphere (unchanged from Lowell), cold but tolerable surface temperatures, and seasonal changes probably due to vegetation. Concerning the polar caps, de Vaucouleurs wrote: "It is hardly necessary to mention that the alternative hypothesis of carbon dioxide snow (dry ice) is no longer tenable" (ref. 9, p. 193). "Infrared reflectance spectra of the polar caps show conclusively that they are not composed of frozen carbon dioxide" is how the Space Science Board put it (ref. 8, p. 86). There was even discussion of martian nuclear weapons tests as the cause of reported flashes¹¹.

Although belief in the reality of the canals began to fade even before the turn of the century, with the last credible sightings by Trumpler in 1924 and Fournier and Pettit in 1939, confidence in a vegetative origin for the "wave of darkening" seems to have reached its peak at the dawn of the space age. E. C. Slipher, writing in 1964, was direct: "Not a single thing has been detected that it does not explain. Every year adds to the number of those who have seen the evidence for themselves. Thus theory and observations coincide." The Space Science Board concurred: "The evidence taken as a whole is suggestive of life on Mars. In particular, the response to the availability of water vapour is just what is to be expected of a planet now relatively arid, but which once probably had much more surface water" (ref. 8, p. 90). More measured, but no less Lowellian, was de Vaucouleurs' summary: "we might liken [the physical conditions on Mars] to those which obtain on a terrestrial desert, shifted to the polar regions and lifted to stratospheric level. We leave it to the reader to decide whether under such circumstances Mars can be 'the Abode of Life' or not" (ref. 9, p. 41). Alternatives to vegetation, in particular windblown dust, were offered but seem to have been unpopular (for example, de Vaucouleurs (ref. 9, p. 269) dismisses "the non-vegetative hypothesis" with a few sentences of exposition and two pages of objections), for much the same sort of reasons that Lowell gave for preferring water over frozen carbonic acid gas.

Box 2

Lowell and Wallace

Alfred Russel Wallace, the co-discoverer of evolution by natural selection, was aged 83 in 1907 when he issued his classic response *Is Mars Habitable?*¹⁵ Wallace began by accepting the canals as a debating point; he had a deeper purpose. Lowell's arguments were generally much better than Wallace makes them out to be, which is why many survived substantially intact into the 1960s, but Wallace identified three fatal flaws, two in physics, the third in interpretation. In order to justify liquid water and clement climates on Mars, Lowell was forced to argue that Earth's albedo is 75%. Thus Mars, with an albedo of 27%, would be comfortable. Wallace had little doubt that this high albedo was wrong — clouds have albedos less than 75% and Earth is no more than half cloud-covered.

Nor did Wallace think the colour of the polar collar a strong argument; on the contrary, he thought the argument ridiculously weak. He suspected the meltwater, were it water, to be necessarily shallow, turbid and muddy brown; not blue. Lowell never compared the amount of water in the martian snowfields to the requirements of the canals, but Wallace did. He found that "the water supply is ludicrously inadequate" to justify anything approaching the "over one hundred thousand miles of canals" that Lowell had mapped (ref. 15, p. 26). In the end, Wallace aligned himself with the candidate of uncommon subtlety: the polar caps are dry ice, and temperatures are correspondingly and uncompromisingly low. On these grounds he concluded that Mars is not habitable. With respect to the polar caps, availability of water and surface temperature, Wallace was right. Mars is not in any apparent sense currently habitable.

The deeper argument was with respect to interpretation. Here Wallace answered the "Nothing I can think of apart from life can explain..." or the "I'll know it when I see it" arguments that still pervade the abstract business of life detection, be it in ancient terrestrial rocks, martian meteorites, european oceans or extrasolar planets. All that is required is to supply a possible counter-example. This Wallace does, in two ways. He closes the book by offering a physical model of fracturing of a cooling, originally hot lithosphere accreted around a primordial cold core. But first he cuts to the quick, by illustrating that different minds can draw different conclusions from the same data. He questions the canals as works of art. Why so straight? (Irrigation canals generally follow the contours of the land, after all.) Is there no topography at all on Mars? How is that possible? What sense is there to this dense network? Where is the design? Why more than 100,000 miles of canals to move such tiny amounts of meltwater? And so on. One is almost bound to agree with him that the canals, "as Mr. Lowell describes, would be the work of a body of madmen rather than of intelligent beings".

Yet around 1963, the imminent launch of the first Mars probes seems to have spurred new work, and with the new work, revisionism. New evidence of a much thinner atmosphere led to new interest in dry ice and a strange suggestion that Mars was coloured by, and sterilized by, abundant nitrogen tetra-oxide (see ref. 12 for a journalist's survey of the then current martian climate). In the end, it was the success of Mariner IV in 1965 that destroyed all hope and smashed all dreams. Mariner IV revealed a heavily cratered moonscape where canaliform features ought to have been. Worse, Mariner IV revealed the dismal truth: "The thinness of the martian atmosphere has been one of the great disappointments of the space age" was how Anders and Owen¹³ opened their landmark study of martian volatile inventories. In this suddenly chilled climate, dry ice leapt back to favour as the chief constituent of the seasonal polar caps¹⁴. The sense of disappointment is still bitter after 35 years. The word echoes through the pages of *To Utopia and Back*, Norman Horowitz's excellent history of the Viking programme and its roots in Lowell's

visions. The reality that Mars is probably dead is still so hard to accept that revivification remains the central thrust of NASA's unmanned planetary exploration programme.

The other great distributary of Lowell's legacy flows through popular culture. Martians and canals established an immediate foothold in the public imagination, with illustrations of martian canals even used to sell cigarettes (Fig. 2; and see ref. 5, p. 74). It was only 12 years ago that the Vice-President of the United States referred publicly to the canals on Mars as fact: "We have seen pictures where there are canals, we believe, and water. If there is water, that means there is oxygen. If oxygen, that means we can breathe." Lowell's Mars was also enlisted by the utopians. Mars was held up as an example of global peace and cooperation, by Lowell and others, for how else could such an intricate system of canals be constructed and maintained? In his tongue-in-cheek explanation (to the arch-Republican Lowell) of his poem 'The Gospel from Mars', E. H. Clement wrote: "My main object is to show that such a system as has developed on Mars, the people of all nations must cease regarding boundary lines and have become one brotherhood, in fact and deed [sic] — so likewise the social classes. In short I am going to show why Mars is carrying through our Heavens the heart-red flag of socialism!" (quoted in ref. 6, p. 220).

A visit to the local science-fiction bookseller reveals an Amazon of hopeful Mars lore flowing underground. H. G. Wells' *War of the Worlds*, published in 1898, is perhaps the best known early example. It is my belief that this huge aquifer of fictional information is the main source of new hope for the dead planet. I had meant to research some of this fiction. Preliminary findings based on cover art, beginning with Edgar Rice Burroughs, reveal that martians developed silicone implants and photosynthetic skins before earthlings did. Ray Bradbury's *Martian Chronicles* and Arthur C. Clark's *Sands of Mars*

have been cited as formative by many space scientists; more recently a rainbow of Mars books (Red, Green, Blue and White; other colours still available) have caught attention among the Mars community. Of the arguably martian novels, my own personal favourites are *War of the Worlds* and Frank Herbert's *Dune* (the latter martian only in the sense that it takes place on a profoundly desiccated world that takes its water from the polar caps). What does inspire me are Chesley Bonestell's paintings of unimaginably ancient canals, still greening the martian desert, their Ozymandian builders otherwise long forgotten (Fig. 3).

On today's Mars, the Abode of Life is likeliest to be underground and cryptic, in deep liquid water aquifers that we hope exist. Of course we must still look for oases, but we cannot assume that we will find one. In some ways the debate has really moved little since the days of Flammarion and Lowell. The most interesting information remains right at the limit of resolution, be it metres in satellite images of gullies, or nanometres in microscopic images of magnetite crystals. Always life on Mars seems just beyond the fields that we know.

Things may have been different in the past, as they are likely to differ in the future. A key is space travel. Currently, traffic between Earth and Mars is fitful, accidental and expensive, depending mostly upon launching inadvertent microbionauts amongst a welter of impact ejecta, and then depending upon the vagaries of orbital dynamics to deliver the few survivors intact and alive to their new home. Today's Mars is too hostile to provide much hope that the surviving colonists could establish a viable self-sustaining ecosystem: habitable niches at the surface are likely to be so few, short-lived and widely separated that extinction seems the likeliest option, even were the first rock to land in a pond, while Mars' cryptic lakes would be unreachable, even if they exist.

In the deep past, such accidental traffic would have been much greater than today, and we have direct observational evidence that Mars was once a far more fit place to live. As it is not known why ancient Mars was in some average sense warmer and wetter than it is today, one cannot go much further into reconstructing the ancient island biogeography between planets. A plausible hope is that Mars still preserves on its surface evidence of a life that left Earth (or vice versa) more than 4 billion years ago. We cannot know this until we go.

The future, we can guess. Eventually earthlings will take Mars for their own. If there still remains at this late date indigenous life on Mars, it will be exterminated. No other course seems credible, our best intentions notwithstanding. Thus we might come to wish that Mars be now sterile, lest we add planetocide to our dictionaries. □

"And before we judge of them too harshly, we must remember what ruthless and utter destruction our own species has wrought... The Tasmanians, in spite of their human likeness, were entirely swept out of existence in a war of extermination waged by European immigrants, in the space of fifty years. Are we such apostles of mercy as to complain if the Martians warred in the same spirit?"

H. G. Wells *The War of the Worlds*



If martian nuclear weapons tests were the cause of observed flashes on Mars in the 1960s, perhaps it was in preparation for a future assault on Earth...

1. Lowell, P. *Mars* (Longmans, Green and Co., London, 1896).
2. Flammarion, C. *La Planète Mars et Ses Conditions d'Habitabilité* (Gauthier-Villars, Paris, 1892).
3. Sheehan, W. *The Planet Mars* (Univ. Arizona Press, Tucson, 1996).
4. Lowell, P. *Mars and Its Canals* (Macmillan, London, 1906).
5. Caidin, M. & Barbree, J. *Destination Mars* (Penguin Studio, New York, 1997).
6. Hoyt, W. G. *Lowell and Mars* (Univ. Arizona Press, Tucson, 1976).
7. Lowell, P. *Mars as an Abode of Life* (Macmillan, London, 1908).
8. Horowitz, N. *To Utopia and Back* (Freeman, New York, 1986).
9. de Vaucouleurs, G. *The Physics of the Planet Mars* (Faber and Faber, London, 1954).
10. Kellogg, W. W. & Sagan, C. *The Atmospheres of Mars and Venus* Publication 944 (National Academy of Sciences National Research Council, Washington DC, 1961).
11. Salisbury, F. B. Martian biology. *Science* **136**, 17–26 (1962).
12. Sullivan, W. *We Are Not Alone* (McGraw Hill, New York, 1964).
13. Anders, E. & Owen, T. Mars and Earth: origin and abundance of volatiles. *Science* **198**, 453–465 (1977).
14. Glasstone, S. *The Book of Mars* (National Aeronautics and Space Administration, Washington DC, 1968).
15. Wallace, A. R. *Is Mars Habitable?* (Macmillan, London, 1907).

Acknowledgements

Thanks to J. Moore for many talks and access to the Moore Memorial Library of Mars Arcana.

Mars' core and magnetism

David J. Stevenson

California Institute of Technology, 150-21, Pasadena, California 91125, USA (e-mail: djs@gps.caltech.edu)

The detection of strongly magnetized ancient crust on Mars is one of the most surprising outcomes of recent Mars exploration, and provides important insight about the history and nature of the martian core. The iron-rich core probably formed during the hot accretion of Mars ~4.5 billion years ago and subsequently cooled at a rate dictated by the overlying mantle. A core dynamo operated much like Earth's current dynamo, but was probably limited in duration to several hundred million years. The early demise of the dynamo could have arisen through a change in the cooling rate of the mantle, or even a switch in convective style that led to mantle heating. Presently, Mars probably has a liquid, conductive outer core and might have a solid inner core like Earth.

Although the existence of the martian core has been accepted for many decades, is interesting for several reasons. First, its size and composition tell us about Mars as a whole — its constituents and provenance. Second, its antiquity tells us about early conditions on Mars; we believe that the core formed early, and this requires that Mars had a hot beginning. Third, this core is the likely source of a magnetic field for some part of Mars' history, probably the earliest part, just as Earth's core is the source of the current geomagnetic field. Fourth, the field may have influenced the early climate through its influence on atmospheric escape. It could also have affected the environment for early life on Mars. Fifth, the heat flow from the core may have fed mantle plumes and influenced volcanic activity, much as hot spots such as Hawaii are thought to be fed by core heat flow on Earth. Sixth, a core, if partly or entirely liquid, influences rotational dynamics, just as (for example) changes in length of day are influenced by Earth's liquid core.

Mars is built from roughly the same ingredients as Earth: silicates and oxides of magnesium and iron, as well as metallic iron (alloyed with various constituents). The mantle and crustal components are discussed in the accompanying article by Zuber (pages 220–227). When we refer to a core for the terrestrial planets Mercury, Venus, Earth and Mars, we mean a central region that is rich in metallic iron. Because this material is about twice as dense as the silicates and oxides making up the crust and mantle, its presence as a core is revealed through its influence on the mean density of the planet and through its effect on the moment of inertia. Old measurements of gravity and more recent geodetic data from Pathfinder¹ reveal that the mean moment of inertia for Mars is $0.365MR^2$, where M is the mass of Mars and R is its mean radius. Together with the martian mean density of 3.93 g cm^{-3} , this suggests a model of Mars that is not too different from a scale model of Earth (that is, similar ingredients, distributed similarly). The martian core is proportionately a little smaller than Earth's core, and proportionately more iron is found in the mantle (in oxides or silicates). As we do not know the composition of the core we cannot be certain about core size, but a core radius of around 1,300–1,500 km (depth to core of 1,900–2,100 km) is indicated. Figure 1 shows a simple interior structure of Mars. Bertka and Fei² suggest a composition for Mars that is different from partly devolatilized primitive meteorites.

Far less is known about the martian core than Earth's core because we lack seismological evidence or geodetic data of sufficient precision. In particular, we do not know whether the core is entirely liquid, partly liquid (like Earth) or entirely solid, although there are indirect arguments against an entirely solid core. Fortunately, because Mars is at lower pressures than Earth, and so is more accessible to high-pressure experiment, it is possible to assess the likely phase composition for the core.

I begin with a discussion of the timing of core formation and the new evidence on the nature and origin of martian magnetism. An assessment of the dynamo process follows and is applied to the possible thermal histories of Mars. I conclude with other implications of martian magnetism and core structure and some comments on future exploration.

Core formation

It is widely accepted that terrestrial planetary cores owe their existence to a process of gravitational separation of mostly liquid, immiscible iron from the (partly) solid silicates. The supporting arguments are partly physical³, but increasingly geochemical. Although we have no samples of either the core of Earth or the core of Mars, we do have rocks that are probably indicative of mantle composition. For Mars, these are the very limited yet highly important SNC meteorites (for shergottites, nakhlites and chassignites). As on Earth, these igneous rocks show a striking depletion of 'iron-loving elements' (called siderophiles) whose extraction testifies to the conditions of core formation⁴. Isotopic data^{5,6} also suggest that this core-forming event was early in Mars' history.

These data, together with physical modelling, suggest a scenario similar to the following. Mars accumulated from smaller bodies over a period of perhaps as long as 100 million years (Myr), but possibly much shorter, around 4.5–4.6 billion years (Gyr) ago^{7–9}. Isotopic evidence is compatible with a very short accretion time, suggesting that Mars might even have been a runaway, isolated embryo rather than a slowly accumulated body like Earth. In this accretional process, the impacting bodies may have already had iron cores, but the energetics of the impact events would have caused extensive melting and mixing of the immiscible metallic iron and silicate/oxide components, allowing chemical re-equilibration on a small scale (centimetres to metres)³. A substantial mass of the impacting bodies may have been in the form of giant impacts (bodies of the order of the mass of Earth's moon),

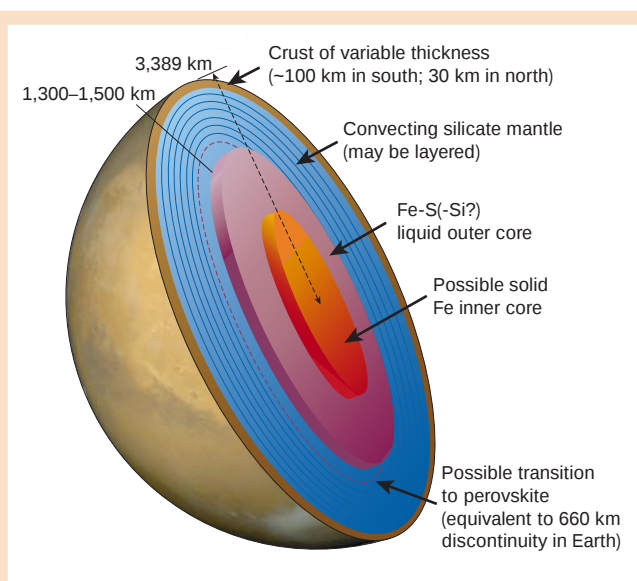


Figure 1 Cutaway view of the martian interior.

but even much smaller bodies can bury a great deal of heat at depth. The energy of gravitational formation of Mars is roughly $0.6GM/R$ per unit mass, where G is the gravitational constant. If all this were converted into heat it would be sufficient to heat the Mars-forming material to several thousand degrees above the melting point, and even with the loss of heat by radiation, a magma ocean is likely. This ocean might be transient (surviving for a brief period after each giant impact), or it might be sustained by a dense steam atmosphere¹⁰, but in either case it will define the conditions of most of the core formation.

Although not as hot or at such high pressure as the likely conditions that formed Earth's core⁴, the lower gravity on Mars would still permit a thick magma ocean. Metallic iron can settle as droplets in a convecting magma ocean, to accumulate as large blobs ('diapirs') that then ascend by Stokes flow through the possibly more viscous and high-pressure deep mantle. In this scenario, the core might initially be either a few hundred degrees hotter than the mantle¹¹ (if the energy of core formation is retained substantially within the iron) or the same temperature as the deep mantle (if efficient thermal equilibration takes place). Much of the martian crust (particularly that preserved in the south) may have formed in this very earliest epoch. Constraints on the timing of crustal formation and thickness are discussed by Zuber (pages 220–227).

Martian magnetism

Mars, unlike Earth, has no global dipole magnetic field. The Mars Global Surveyor spacecraft confirmed this, but also found strong, spatially variable magnetic fields at altitudes of ~200 km down to closest approach of ~110 km (refs 12, 13). Figure 2 shows hemispherical maps of the radial field normalized to a constant 200-km altitude. The fields are measured below much of the martian ionosphere and much of the power in their spatial variability is at length scales comparable to the distance from the surface. 'Inversion' of these data is non-unique, but the source of the field must be confined to the outermost several tens of kilometres of the crust (and possibly confined to an even thinner layer). A deeper layer or source of currents could not provide the observed spatial structure, except with physically implausible assumptions. The inferred crustal magnetizations are up to ~10–30 A m⁻¹, an order of magnitude higher than the strongest magnetizations typically encountered in Earth rocks, and even these values are underestimates if one were to require thinner magnetized layers or incoherent magnetization directions. Given the large amounts of magnetized crust required, it seems very

likely that the magnetization is thermal remanence acquired during the last time the rocks cooled through the blocking temperature, for example, following dike injection¹⁴. Most of these cooling events took place at a time when a large global field was present. ('Most' rather than 'all' is appropriate here, because the surface magnetizations are so large that it is possible for crustal rock to be substantially magnetized through cooling in the presence of other crustal fields, rather than a global field.) Other origins of the magnetization (for example, due to impact, as suggested for the Moon) are conceivable in principle, but seem insufficient given the magnitude of the requirements.

There is great interest in the meaning of the spatial pattern of magnetization, including possible lineations that suggest an analogy to plate tectonically derived lineations of magnetization on Earth's ocean floor, but the current constant-altitude (that is, constant-resolution) maps do not provide strong support for these speculations. Models of the crustal magnetization suggest that the martian field may have undergone reversals.

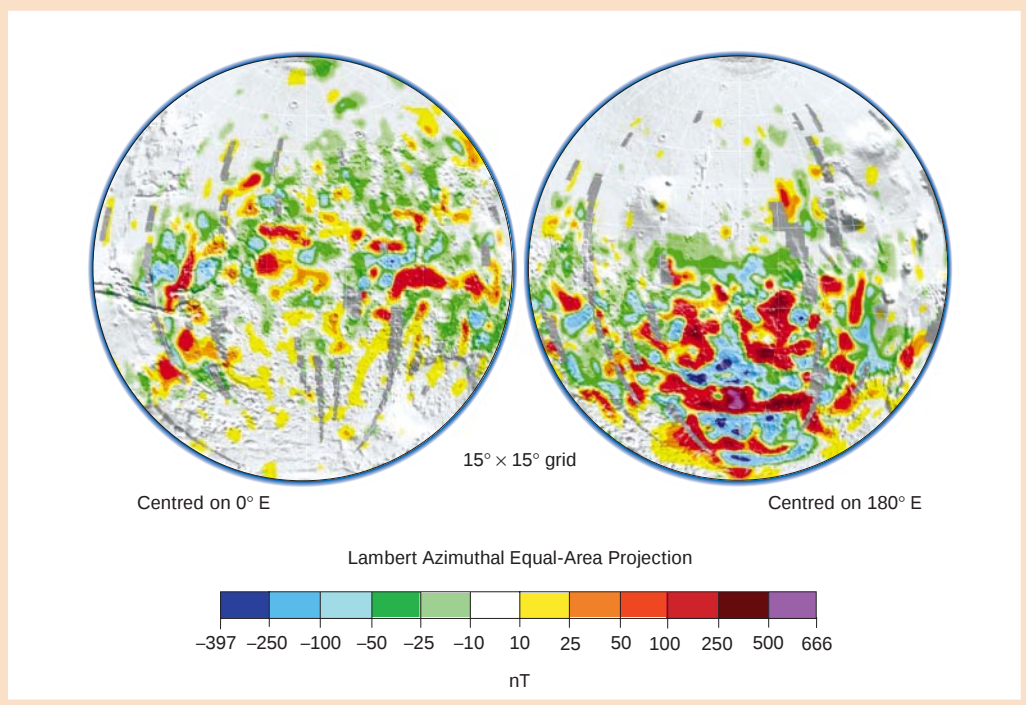
It seems likely that Mars requires at least one (and preferably several) of the following: high abundance of appropriate magnetic materials (for example, magnetite), a particularly favourable magnetic mineralization (for example, single domains), large volumes of crust that are coherently magnetized, and/or an unusually large field in which the magnetization was acquired.

The fact that Mars did have a global magnetic field for one or more periods in its early history suggests that it once had an active core dynamo, the process responsible for Earth's current field. The strongest magnetizations are observed in the ancient southern highlands of Mars, which predate 4 Gyr. The antiquity of these regions is inferred by the (imprecise) method of crater counts (see review in this issue by Zuber, pages 220–227). However, not all ancient crust on Mars produces large magnetic fields at the spacecraft altitude, and not all younger crust is devoid of magnetization. This prevents firm conclusions being made about the timing of acquisition of magnetization and hence the timing of a postulated martian core dynamo.

Schubert *et al.*¹⁵ have suggested a later (post ~4.0 Gyr) period of magnetization, which would indicate a later period of dynamo activity. But some arguments point towards ancient (4.2 Gyr or earlier) acquisition. First, as noted by Acuna *et al.*¹², the ancient impact structure Hellas (believed to be at least 4 Gyr old) lacks any magnetic signature, and seems to be surrounded by a region with very little coherent magnetization. This is a plausible outcome were the impact to have occurred when Mars possessed no global magnetic field. It is not a plausible outcome if the southern crust were subsequently reheated and then cooled to acquire magnetization during a later epoch in which a global field was active. Second, it is difficult to imagine any physically plausible scenario in which large provinces of the southern crust were extensively heated later in Mars' history without producing some surficial difference in appearance from those regions that were not so treated. This argument is supported by the recognition that huge volumes of crust are required to explain the observed magnetization, rather than some thin layer of possibly remagnetized material. In particular, revived igneous activity generating new crust would certainly disrupt these terrains because of the large amounts required. Third, evidence from the ancient martian meteorite ALH84001 suggests that its magnetization was acquired at 4.0 Gyr or even earlier¹⁶.

A region that lacks large magnetic fields at the spacecraft altitude might still consist of crust that formed in the presence of a global magnetic field. For example, the magnetization may be spatially incoherent, the cooling history may have favoured multidomain magnetite or less favourable mineralization, or the field may have been reversing more rapidly. Moreover, the early rapid pace of planetary evolution means that regions in the south that seem to be of the same age may nonetheless differ in age by ~100 Myr and thus cooled in a different magnetic field, even though their surface

Figure 2 Radial magnetic field at 200-km altitude, based on data collected by the Mars Global Surveyor spacecraft. The southern hemisphere exhibits the largest field anomalies. Notice that the opposite hemisphere has field anomalies that are typically an order of magnitude smaller and exhibit less coherent spatial structure. These maps were prepared by M. Purucker, Goddard Space Flight Center. They are in Lambert Azimuthal Equal-Distant projection and were published (in a different projection) in ref. 39.



appearance and crater density seem to be identical. Last but not least, ancient crust may underlay younger crust in some northern localities, thereby allowing preservation of a (relatively weak and patchy) magnetization, even when the surface age postdates any global field.

A martian dynamo?

The dynamo mechanism (see Box 1) is much studied but still imperfectly understood^{17,18}, despite recent advances in numerical simulation¹⁹. In particular, we do not know the conditions sufficient for the existence of a planetary dynamo. Because we can only speculate about early conditions on Mars, the problem of inferring or predicting the history of a martian dynamo is indeed formidable. Earth's dynamo is also imperfectly understood, although it is thought that it arises from convection driven largely by inner-core growth^{20–22}.

If a dynamo exists, then it is likely that the expected field magnitude B inside the region of field generation is given by the Elsasser number of order unity. This implies $B \sim (2\rho\Omega/\sigma)^{1/2}$, where ρ is the fluid density, Ω is the planetary rotation rate and σ is the electrical conductivity. For the martian core, this yields $\sim 10^{-3}$ tesla, but because this is the same prediction as for present Earth, the predicted palaeofield at the surface of Mars is indistinguishable (at the level of this crude argument) from the present field at Earth's surface. Dynamo theory admits weaker fields as possible solutions, but it does not admit fields substantially larger than $B \sim (2\rho\Omega/\sigma)^{1/2}$. Despite suggestions to the contrary²², there is no theoretical basis at present for the idea that the field scales in some direct way with the energy source, so that it might undergo slow decline over geological time or large changes arising from inner-core nucleation.

One speculative explanation for the origin of magnetization on Mars is that the field was generated in a magma ocean. Plausible numbers are a characteristic fluid velocity v of $\sim 10^{-1} \text{ m s}^{-1}$ (because of very high heat flows at that time), a characteristic length scale L of $\sim 10^6 \text{ m}$, and a magnetic diffusivity λ of $10^4 \text{ m}^2 \text{ s}^{-1}$ (possibly appropriate to high-temperature and high-pressure silicate melts²³), which together give a magnetic Reynolds number R_m of ~ 10 . This is marginal at best, but would be attractive because large fields are predicted (~ 0.01 – 0.1 tesla at the martian surface). The

extremely high observed magnetizations might then be explained, although a core dynamo is more plausible.

Thermal or compositional convection?

If one accepts that core convection is needed, then a probable necessary condition for a dynamo is the presence of convection. In terrestrial planets (including Earth), the criterion for core convection is difficult to satisfy. The reason for this is that the natural scale for core heat flows is such that this heat can probably be carried by conduction at a temperature gradient that is stably stratified (that is, it inhibits convection). To obtain core convection, one must appeal to unusually large heat flows or the development of an inner core. In either case, the core must be cooling. To appreciate this argument, consider first the simple case of no inner core. Convection will occur provided the heat flux within the core exceeds that which is carried by conduction along an adiabat:

$$F_{\text{total}} > F_{\text{cond,ad}} \equiv k\alpha Tg/C_p \Leftrightarrow \text{thermal convection}$$

where k is the thermal conductivity, α is the coefficient of thermal expansion, T is the temperature, g is the gravitational acceleration and C_p is the specific heat at constant pressure. These parameters are all slowly varying within a core (if T is close to being adiabatically distributed), except for g , which is approximately linear in radius r , the distance from the planet centre. If the core is simply cooling and releasing the stored sensible heat (provided by gravity during planetary accretion), then the total heat flux is also linear in r : $F_{\text{total}}(r) = -\rho C_p r (dT_c/dt)/3$, where T_c is the mean core temperature and t is time. It is unlikely that the core contains significant radioactive heat sources (even less likely than Earth, where one can always appeal to unknown, very high pressure effects). Consequently, if thermal convection ceases to operate in the outer part of the core, then it will also cease to operate at about the same time elsewhere in the core.

If the core is cooling and the central temperature drops below the liquidus for the core alloy, then an inner core will nucleate. In Earth, we know from seismic evidence that the core is $\sim 10\%$ less dense than pure iron, and many suggestions have been offered for the identity of the light elements that are mixed with the iron^{24,25}. At the lower pressures relevant to Mars, the dominant light element may be sulphur². For sulphur abundances that are less than cosmic relative to iron, as cosmo-

chemical arguments suggest, the inner core will be nearly pure iron (with some nickel) and the sulphur will be entirely in the outer core. The introduction of this light element into the fluid of the lowermost core will tend to promote convection and cause mixing throughout all or most of the outer core, provided the cooling is sufficiently fast. Latent heat release at the inner-core/outer-core boundary will also contribute to the likelihood of convection. However, inner-core growth permits outer-core convection even when the heat flow through the core–mantle boundary is less (perhaps much less) than the heat carried by conduction along an adiabat. In this regime (possibly that occupied by Earth), the temperature gradient is very slightly less steep than adiabatic and the compositional convection carries heat downwards. The total heat flux is still outwards, of course, as the heat carried by conduction is large. This state is possible because the buoyancy release associated with the compositional change exceeds the work done against the unfavourable thermal stratification. Unlike thermal convection, compositional convection may not cease everywhere throughout the core at a single epoch. This argument is modified in detail but not in general outline, should the core include a light element that does not exhibit eutectic behaviour (for example, silicon), as well as the (universally expected) complement of sulphur.

Required cooling rates

For plausible choices of parameters, the cooling rate of the core must exceed about 80 K Gyr^{-1} to obtain thermal convection. However, this estimate is uncertain by perhaps as much as a factor of two. The required cooling rate in the presence of a growing inner core is much smaller, by a factor of several^{26,27}, but has not been studied in detail for Mars. As a consequence, models with an inner core will tend to sustain a dynamo for a long time, perhaps even to the present day, unless there is something unusual about the thermal history (as suggested below). The overlying mantle determines the cooling rate. Indeed, it is the mantle that determines whether a terrestrial planet has core convection and whether it can have a dynamo.

It is also possible that the sufficient condition for a dynamo is not far removed from the necessary condition for the presence of any convection²⁶. Simple scaling laws for convection (compatible with the philosophy of Kolmogorov turbulence and known to astrophysicists as mixing length theory) suggest that $v \sim 0.1(F_{\text{conv}}/\rho)^{1/3}$, where ρ is the fluid density and F_{conv} is the convective heat flux (or its compositional equivalent when the convection is driven by compositional density differences). I define

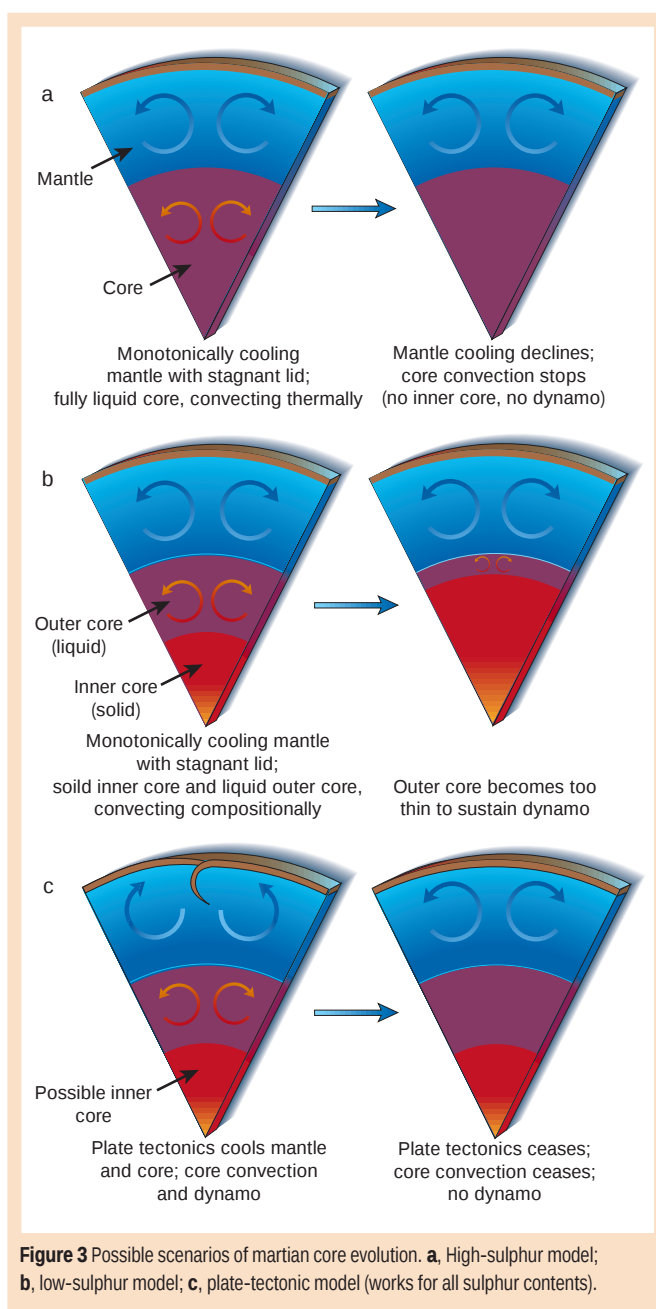
$$\varepsilon = F_{\text{conv}}/F_{\text{cond,ad}} = (F_{\text{total}} - F_{\text{cond,ad}})/F_{\text{cond,ad}}$$

Substitution above shows immediately that for plausible parameters in the martian core ($L \sim 10^6 \text{ m}$, $\lambda \sim 1 \text{ m}^2 \text{ s}^{-1}$, $\rho \sim 10^4 \text{ kg m}^{-3}$), R_m may be large even if $\varepsilon \ll 1$. That is, the heat flow has to only slightly exceed that for any convection in order to reach that for convection of sufficient vigour to sustain a dynamo. This claim must be tested by further numerical work. It is conceivable, but difficult energetically, for a dynamo to function for $\varepsilon < 0$ (for example, because of baroclinic instabilities and thermal winds arising from horizontal temperature gradients that are caused by lateral differences in heat flow through the core–mantle boundary). Even in this case, one would expect that a dynamo requires $|\varepsilon| \ll 1$, as the vertical motions would otherwise be strongly suppressed and this inhibits dynamo activity. Alternatives to convective driving (for example, precession²⁸) still require the core to be close to adiabatic and thus do not escape the constraints discussed above.

In conclusion, if the mantle cools fast enough (or is cool enough to allow inner-core nucleation) then a dynamo occurs, but if the mantle is too hot or fails to cool then there is no dynamo.

Possible histories of the martian core

Three possible scenarios for the history of the martian dynamo are presented in Figure 3. The first is the simplest: the planet starts out very hot and cools quickly at first. The core remains completely liquid



throughout. As the cooling rate declines, a point is reached at which the heat flow out of the core can be accommodated by conduction alone. At that epoch, the dynamo turns off (in a very short time geologically, perhaps as little as a few thousand years) and no further field generation is possible, provided an inner core never develops. This model requires that the core of Mars is sulphur-rich, perhaps 10% or more by mass. It also requires tuning of the parameters so that the dynamo turns off as early in Mars' history as some arguments suggest. Most published models are of this kind^{22,29}.

The second scenario is almost the antithesis of the first. It has not been modelled in detail, although it is implicit in the early work of Young and Schubert³⁰, who considered the possibility of complete core freezing. In this model, the sulphur content for the core is sufficiently low that an inner core develops early and grows rapidly. The liquid outer core becomes progressively more sulphur-rich and evolves towards the eutectic composition. The experimental data³¹ yield a eutectic of around 1,400 K at the top of the core. For realistic models of Mars' mantle convection³², the expected present-day core–mantle boundary temperature is at least 1,850 K, so there is no

Box 1

What is a dynamo?

The essence of a dynamo lies in electromagnetic induction — the creation of currents and associated field through the motion of conducting fluid across magnetic field lines. Numerical and analytical work indicate that a dynamo will exist if the fluid motions have certain desired features and the magnetic Reynolds number R_m exceeds about 10. Here,

$$R_m \equiv vL/\lambda$$

$$\lambda \equiv 1/\mu_0\sigma$$

where v is a characteristic fluid velocity, L is a characteristic length scale of the motions or field (for example, the core radius), λ is called the magnetic diffusivity, μ_0 is the permeability of free space and σ is the electrical conductivity (SI units). It seems likely that fluid motions of the desired character arise naturally in a convecting fluid (irrespective of the source of fluid buoyancy), provided the Coriolis force has a large effect on the flow, that is, $v/\Omega L < 1$, where Ω is the planetary rotation rate. This is easily satisfied for any plausible fluid motion of interest. Realistic dynamos often require a somewhat larger value than ~ 10 for R_m , especially if driven by secular cooling without an inner core⁴⁰, and attention must also be paid to the different kinds of motions of relevance (for example, vertical motions and differential rotation may have different amplitudes and scales of variation).

prospect that the outer core will completely freeze. However, it is conceivable that the outer core will become sufficiently thin that dynamo activity can no longer be sustained. This would seem implausible based on simple scaling arguments, but it might be the state that Mercury currently occupies³³. It probably requires a lower sulphur content of the martian core than most would consider plausible, perhaps no more than a few per cent (even less than typical estimates for Earth). Further dynamo simulations are needed to test this hypothesis.

The third scenario invokes a change in mantle convection to trigger the death of the martian dynamo³². It is assumed that early Mars had mobile lid convection in which the lithosphere could be recycled. On Earth, this is accomplished by plate tectonics, and this could also be the case on Mars³⁴. (It is, however, the recycling of the lithosphere that matters, not the form of the recycling; so there is no need to assume that Mars did exactly what Earth does.) At some time, perhaps after only a few hundred million years, this process ceased and Mars evolved slowly into the stagnant lid regime that it (and all terrestrial bodies except Earth) currently occupies. If this regime follows one of lithospheric recycling, then the mantle must heat up, because the elimination of heat is less efficient. In other words, the coldest time for the martian mantle was early in Mars' history, despite the inexorable monotonic decline of radioactive heat sources in the mantle and the crust. This scenario has the advantage that it may work for all possible sulphur contents in the core, as the presence of an inner core will not drive a dynamo if the mantle minimum temperature was reached early in Mars' history. An inner core drives a dynamo only while it is growing, and it can grow only if the core is cooling. One problem with this scenario is that it invokes an *ad hoc* timing for the cessation of 'plate tectonics'; it also implies the ability for Mars to be volcanically active throughout geological time.

In all these scenarios, the beginning of dynamo activity may be delayed after Mars' accretion until a thermal boundary layer builds up in the lowermost mantle, depending on the uncertainty in the initial temperature difference between the core and mantle. But this is unlikely to produce a delay of more than ~ 100 Myr.

There may be other scenarios not yet considered. Unfortunately, none of these scenarios can be tested with great confidence because

the parameters that define their chronologies are not known with sufficient accuracy. However, the presence or absence or size of the inner core is clearly a crucial variable and may eventually be determined by a combination of geodesy and seismology. Numerical dynamo modelling will also be important in the coming years.

Consequences of the martian core and dynamo

Core cooling dictates the presence of a thermal boundary layer at the base of the overlying mantle. Plumes can detach from this layer and may be a cause of hot-spot volcanism. Harder and Christensen³⁵ have proposed that Mars may be in a regime where a single plume dominates because of the effect of a major endothermic phase transition near the base of the martian mantle (the same phase transition that defines the upper-mantle/lower-mantle boundary on Earth). This plume might be stable for a long period of time, perhaps billions of years, and may be responsible for the Tharsis volcanic province. This hypothesis provides the exciting prospect of linking core thermal history with martian volcanic history. However, it leaves unanswered several questions. If the core heat flow is so low (as required by the absence of a dynamo throughout much of Mars' history), then is it reasonable to suppose that it is responsible for the dominant volcanic activity on Mars? Why would a deep-seated plume happen to produce volcanism at a location just northward of the principal geological feature (the crustal dichotomy)? Why is the plume so stable? Perhaps the answer to Tharsis lies nearer the surface of Mars rather than in the core history.

The history of the atmosphere³⁶ may also be influenced by the magnetic field history through the effect of the field on atmospheric sputtering. The history of martian magnetism might even be linked to the history of life on Mars. Perhaps the strongest argument for a biological effect in ALH84001 lies in the single-domain magnetite grains³⁷, whose presence in biological organisms is useful only while Mars has a field. This might also push martian magnetism back to the earliest epoch.

The future

Although martian core studies can benefit from work in all areas of planetary science (including geochemistry), the greatest contribution is likely to arise from seismological and geodetic efforts. In particular, the Mars Netlander mission³⁸ and subsequent follow-ups are likely to have the greatest role. It may also be essential to better characterize the surface magnetization, something that no currently funded mission can do. We can also look forward to exciting developments in our understanding of dynamos. Mars' core is at least as interesting as Earth's core for our general understanding of planet evolution. □

1. Folkner, W. N., Yoder, C. F., Yuan, D. N., Standish, E. M. & Preston, R. A. Interior structure and seasonal mass redistribution of Mars from radio tracking of Mars Pathfinder. *Science* **278**, 1749–1752 (1997).
2. Bertka, C. M. & Fei, Y. W. Implications of Mars Pathfinder data for the accretion history of the terrestrial planets. *Science* **281**, 1838–1840 (1998).
3. Stevenson, D. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. E.) 231–250 (Oxford Univ. Press, New York, 1990).
4. Righter, K., Hervig, R. L. & Kring, D. A. Accretion and core formation on Mars: molybdenum contents of melt inclusion glasses in three SNC meteorites. *Geochim. Cosmochim. Acta* **62**, 2167–2177 (1998).
5. Chen, J. H. & Wasserburg, G. J. Formation ages and evolution of Shergotty and its parent planet from U-Th-Pb systematics. *Geochim. Cosmochim. Acta* **50**, 955–968 (1986).
6. Lee, D. C. & Halliday, A. N. Core formation on Mars and differentiated asteroids. *Nature* **388**, 854–857 (1997).
7. Wetherill, G. W. Provenance of the terrestrial planets. *Geochim. Cosmochim. Acta* **58**, 4513–4520 (1994).
8. Chambers, J. E. & Wetherill, G. W. Making the terrestrial planets: N-body integrations of planetary embryos in three dimensions. *Icarus* **136**, 304–327 (1998).
9. Agnor, C. B., Canup, R. M. & Levison, H. On the character and consequences of large impacts in the late stage of terrestrial planet formation. *Icarus* **142**, 219–237 (1999).
10. Matsui, T. & Abe, Y. Formation of a magma ocean on the terrestrial planets due to the blanketing effect of an impact-induced atmosphere. *Earth Moon Planets* **34**, 223–230 (1986).
11. Flasar, F. M. & Birch, F. Energetics of core formation: a correction. *J. Geophys. Res.* **78**, 6101–6103 (1973).
12. Acuna, M. H. et al. Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. *Science* **284**, 790–793 (1999).
13. Connerney, J. E. P. et al. Magnetic lineations in the ancient crust of Mars. *Science* **284**, 794–798 (1999).
14. Nimmo, F. Dike intrusion as a possible cause of linear Martian magnetic anomalies. *Geology* **28**, 391–394 (2000).

15. Schubert, G., Russell, C. T. & Moore, W. B. Timing of the martian dynamo. *Nature* **408**, 666–667 (2000).
16. Weiss, B. *et al.* Records of an ancient martian magnetic field in ALH84001. *Nature* (submitted).
17. Merrill, R. T., McElhinney, M. W. & McFadden, P. L. *The Magnetic field of the Earth* (Academic, New York, 1998).
18. Busse, F. H. Homogeneous dynamos in planetary cores and in the laboratory. *Annu. Rev. Fluid Mech.* **32**, 383–408 (2000).
19. Roberts, P. H. & Glatzmaier, G. A. Geodynamo theory and simulations. *Rev. Mod. Phys.* **72**, 1081–1123 (2000).
20. Loper, D. E. Some thermal consequences of a gravitationally powered dynamo. *J. Geophys. Res.* **83**, 5961–5970 (1978).
21. Gubbins, D., Masters, T. G. & Jacobs, J. A. Thermal evolution of the Earth's core. *Geophys. J. R. Astron. Soc.* **59**, 57–99 (1979).
22. Stevenson, D. J., Spohn, T. & Schubert, G. Magnetism and thermal evolution of the terrestrial planets. *Icarus* **54**, 466–489 (1983).
23. Tyburczy, J. A. & Fiesler, D. K. in *Mineral Physics and Crystallography. A Handbook of Physical Constants* (ed. Ahrens, T. J.) 185–208 (Am. Geophys. Union, 1995).
24. Li, J. & Agee, C. B. Element partitioning constraints on the light element composition of the Earth's core. *Geophys. Res. Lett.* **28**, 81–84 (2001).
25. Gessmann, C. K., Wood, B. J., Rubie, D. C. & Kilburn, M. R. Solubility of silicon in liquid metal at high pressure: implications for the composition of the Earth's core. *Earth Planet. Sci. Lett.* **184**, 367–376 (2001).
26. Stevenson, D. J. Planetary magnetic fields. *Rep. Prog. Phys.* **46**, 555–620 (1983).
27. Lister, J. R. & Buffett, B. A. The strength and efficiency of thermal and compositional convection in the geodynamo. *Phys. Earth Planet. Int.* **91**, 17–30 (1995).
28. Stevenson, D. J. Planetary magnetism. *Icarus* **22**, 403–415 (1974).
29. Schubert, G. *et al.* in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder C. W. & Matthews, M. S.) 147–183 (Univ. Arizona Press, Tucson, 1992).
30. Young, R. E. & Schubert, G. Temperatures inside Mars: is the core liquid or solid? *Geophys. Res. Lett.* **1**, 157–159 (1974).
31. Fei, Y. W., Bertka, C. W. & Finger, L. W. High-pressure iron-sulfur compound, Fe₃S₂, and melting relations in the Fe-FeS system. *Science* **275**, 1621–1623 (1997).
32. Nimmo, F. & Stevenson, D. J. Influence of early plate tectonics on the thermal evolution and magnetic field of Mars. *J. Geophys. Res.* **105**, 11969–11979 (2000).
33. Schubert, G., Ross, M. N., Stevenson, D. J. & Spohn, T. in *Mercury* (eds Chapman, C. *et al.*) 429–460 (Univ. Arizona Press, Tucson, 1988).
34. Sleep, N. H. Martian plate tectonics. *J. Geophys. Res.* **99**, 5639–5655 (1994).
35. Harder, H. & Christensen, U. R. A one-plume model of martian mantle convection. *Nature* **380**, 507–509 (1996).
36. Brain, D. A. & Jakosky, B. M. Atmospheric loss since the onset of the Martian geologic record: combined role of impact erosion and sputtering. *J. Geophys. Res.* **103**, 22689–22694 (1998).
37. Thomas-Keprta, K. L. *et al.* Truncated hexa-octahedral magnetite crystals in ALH84001: presumptive biosignatures. *Proc. Natl Acad. Sci. USA* **98**, 2164–2169 (2001).
38. Special Issue. *Planet. Space Sci.* **48**, 1143–1420 (2000).
39. Purucker, M. *et al.* An altitude-normalized magnetic map of Mars and its interpretation. *Geophys. Res. Lett.* **27**, 2449–2452 (2000).
40. Kutzner, C. & Christensen, U. Effects of driving mechanisms in geodynamo models. *Geophys. Res. Lett.* **27**, 29–32 (2000).

The crust and mantle of Mars

Maria T. Zuber

Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA
(e-mail: zuber@mit.edu)

Clues to the history of Mars are recorded in the chemistry and structure of the planet's crust and mantle. The mantle is the rocky, interior region of the planet that transports heat generated during accretion and subsequent core formation. The crust formed by melting of the upper mantle, and has been shaped and re-distributed by impact, volcanism, mantle flow and erosion. Observations point to a dynamically active interior in the early phases of martian history, followed by a rapid fall-off in heat transport that significantly influenced the geological, geophysical and geochemical evolution of the planet, including the history of water and climate.

The different geological evolution of Mars compared to Earth is due primarily to Mars' smaller size. The radius of Mars is about half that of Earth, and so Mars probably heated up and cooled off more quickly. Thus, geological activity manifest as tectonism, volcanism and the associated release of volatiles should have occurred relatively earlier for Mars than for Earth. On Earth, convective cooling of the interior drives the motions of surface plates and is accompanied by the creation of seafloor at mid-ocean ridges and consumption of plates at subduction zones. In contrast, Mars is currently a single-plate planet with a thick, rigid outer shell. But it is possible that earlier in martian history, when internal heat loss was more intense, the planet displayed thinner and possibly even mobile plates¹.

Understanding the evolution of the crust and mantle of Mars has been aided significantly by orbital global geophysical measurements of topography, gravity and magnetics, Earth-based and orbital spectra, orbital- and lander-scale images of the surface, and geochemical measurements of martian meteorites. These data collectively tell the story of an early, dynamic martian interior.

Composition of crust and mantle

Crustal composition

Information on the composition of the crust of Mars has been obtained from Earth-based² and orbital spectra³⁻⁶, *in situ* spectral and chemical observations from the Viking and Pathfinder landers⁷, and geochemical analyses of meteorites believed to have come from Mars^{8,9}. The surface is composed of a mixture of relatively pristine igneous rocks overlain by highly oxidized weathering products that constitute the relatively bright dust and soils. The reddish colour of the martian surface is due to the presence of ferric iron-bearing minerals in the oxidized surface layer^{10,11}. At the local scale, the Viking landers revealed the likely presence of peroxide, a reactive oxidant. The surface sampled *in situ* lacks organic compounds, which, although periodically replenished by meteorites, are probably destroyed by the oxidants as well as intense ultraviolet radiation. If early Mars once had a significantly denser atmosphere than present (ref. 12, and see review in this issue by Jakosky and Phillips, pages 237–244), then evidence should be preserved in the crustal surface layer as carbonates and sulphates. Limited amounts of both have been found in martian meteorites⁹, and sulphates have been found *in situ*¹³, but neither has been detected

conclusively from remote spectral data, which are sensitive at levels of 5–10% abundance^{14,15}.

Thermal-emission spectrometer data from the Mars Global Surveyor (MGS) orbiter suggest that dark regions exhibit two compositions⁶: a basalt-rich component in the southern highlands and andesite-rich component in the northern lowlands. On Earth, basalt is a common volcanic rock formed by melting of the upper mantle, while andesite is found almost exclusively in subduction zone environments, where water has been important in the melt generation process. Andesite formed by fractional crystallization of dry basaltic magma would require large amounts of melting⁷, but for water-rich magmas a greater proportion of andesitic magma can be produced^{16,17}. An alternative interpretation is that the proposed andesite spectral features reflect a basaltic composition with a significant glass component¹⁸.

Certain classes of basaltic achondrite meteorites — shergottites, nakhlites and chassignites (collectively termed SNCs) — are believed to have been ejected from the martian surface by one or more impacts¹⁹ within the past 1–20 million years (Myr). Evidence for a martian origin of these meteorites includes relatively young crystallization ages (0.15 to 1.3 billion years (Gyr)) by terrestrial planet standards¹⁹⁻²¹, and the presence of trapped gases that match isotopically the martian atmosphere²²⁻²⁵. Geochemical analyses indicate that all the martian meteorites are picritic or basaltic lavas or were derived from basaltic magmas⁷, which is consistent with visual and near-infrared spectral observations⁵, but not thermal-infrared orbital

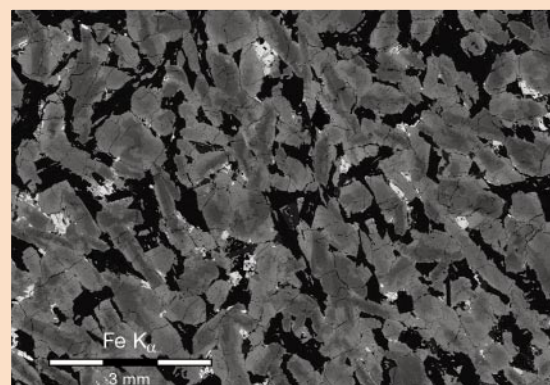


Figure 1 Photograph of a microscopic thin section of the Shergotty meteorite, which is believed to represent a sample of the martian crust.

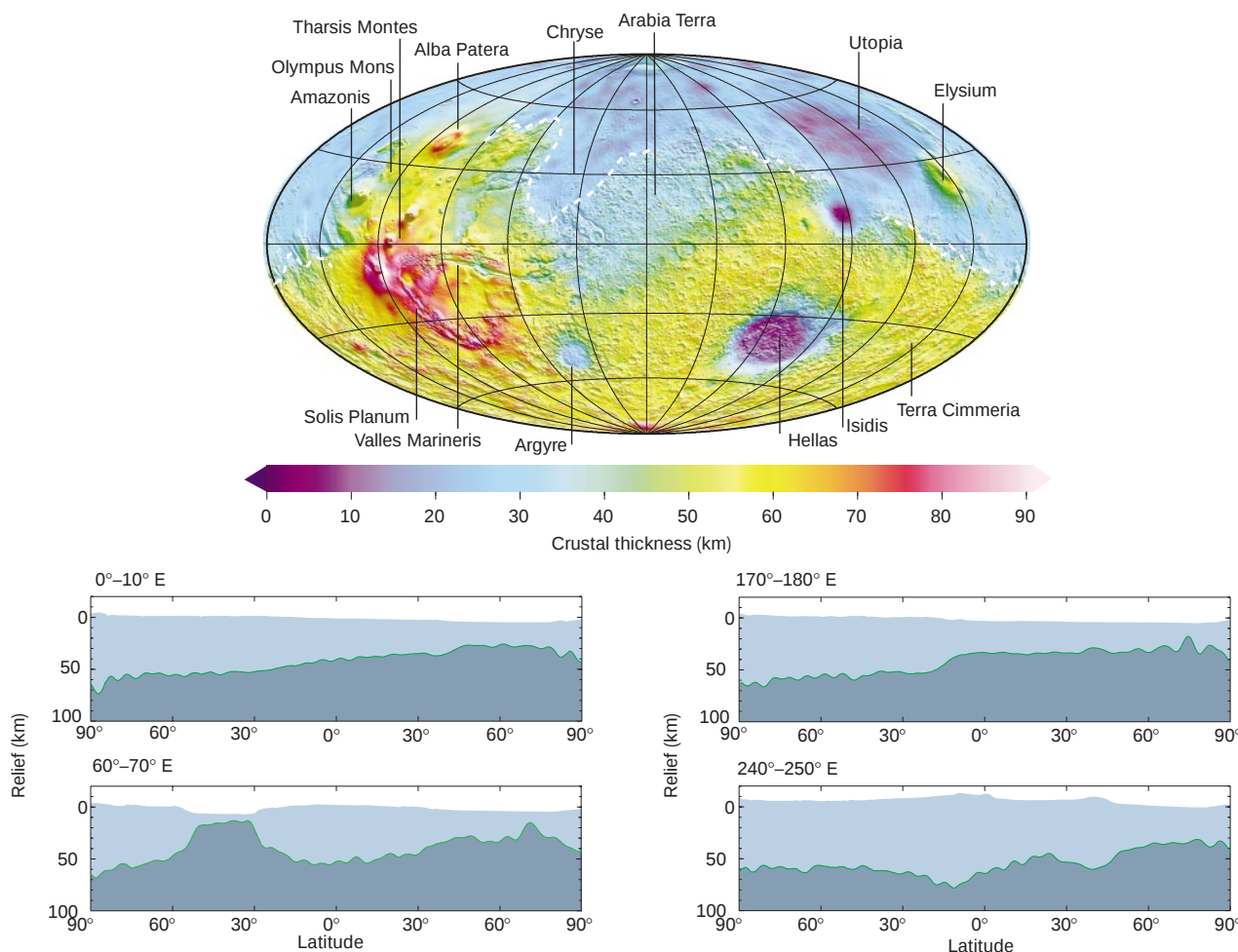


Figure 2 The crustal thickness of Mars assuming a constant-density crust and mantle. The model (crust1004), shown over a shaded relief map, is based on Mars Global Surveyor gravity⁵⁵ and topography⁵⁴ fields. The crustal thickness model has a spatial resolution of ~180 km. The map projection is Mollweide and the coordinate system is areocentric with an east-positive longitude convention. The dashed line shows the location of the geological dichotomy boundary between the northern and southern hemispheres. The boundary is plotted only where it is distinctively expressed. Note that the crustal provinces of the northern and southern hemisphere follow the yellow-green interface and do not correlate everywhere with the surficial expression of the geological dichotomy. Also shown are 10°-averaged south pole-to-north pole longitudinal transects of crustal structure, where light grey corresponds to crust and dark grey corresponds to mantle. The transects for longitudes 0°–10° E and 170°–180° E show crustal provinces of the northern and southern hemisphere. The transect through 60°–70° E shows crustal thinning beneath the Hellas basin, and that through 250° E shows crustal thickening beneath Tharsis.

observations^{6,15}. The parent magmas of the martian meteorites were probably generated by partial melting of the uppermost mantle of Mars and were emplaced at the surface and within the crust by volcanic and magmatic processes. The meteorite ALH84001, renowned for proposed evidence of past biological activity²⁶, has been traced to a martian origin by microscopic-scale textural features and its oxygen isotopic signature²⁷. However, this meteorite has a crystallization age of ~4.5 Gyr and probably represents a sample of ancient crust^{28–30}. Of the martian meteorites, the shergottites (Fig. 1) represent late-stage magmatic products that are considered the most representative samples of unmodified martian crust, but the parent rocks of all martian meteorites sample only a small part of the martian surface.

Mantle composition

Compositions of martian meteorites in combination with models of the planet's density distribution with depth inferred from geophysical constraints^{31–33} lead to inferences about the composition of the martian mantle. It has been generally assumed that the bulk composition of Mars is 'chondritic', that is, approximating the composition of carbonaceous chondrite meteorites^{34,35}, which are believed to be

representative of the most primitive material in the Solar System. It has been suggested that the bulk composition of Mars may deviate from carbonaceous chondrites³⁶ and, unlike the crust, might be characterized by reducing conditions³⁷. A model of the martian interior based on martian meteorites³⁸ suggests that the planet was accreted from two chemically distinct components, the first of which was volatile-rich and oxidizing, and the second was reducing and consisted of high-temperature minerals. These components may have mixed to provide a mantle that is now chemically homogeneous³⁴.

Various bulk composition models of Mars have been proposed and a common feature is enrichment in iron relative to Earth's mantle. Such models can be converted to pressure- and temperature-dependent mineralogies. One plausible structure³⁵, based on the model of Dreibus and Wanke³⁴, has the upper martian mantle similar to Earth's, consisting primarily of the mineral olivine [(Mg,Fe)₂SiO₄]. With increasing depth a transition zone would consist of the more densely packed spinel structure (the high-pressure polymorph of olivine), and the lower mantle would contain a narrow zone rich in an even denser perovskite structure [(Mg,Fe)SiO₃]. Transitions to denser phases occur at greater depth in the martian

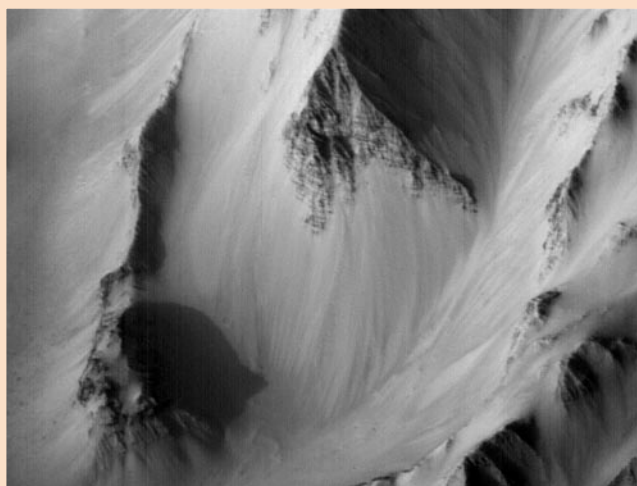


Figure 3 Mars Orbiter Camera image of layering in the western Tithonium Chasma/lus Chasma region of the Valles Marineris, centred at 6.6° S, 269.6° E (ref. 64). The resolution is 6.45 m pixel⁻¹ across by 9.65 m pixel⁻¹ down.

mantle than on Earth because of the lower gravity. For example, the olivine–spinel transition that occurs at 400-km depth on Earth is expected at a depth of about 1,000 km on Mars. Whether or not the spinel–perovskite phase change actually occurs within the mantle depends on the mantle temperature and the size of the core. The moment of inertia factor determined from tracking the Pathfinder lander³³, in combination with estimates of core composition (ref. 39, and see review in this issue by Stevenson, pages 214–219), put the core–mantle boundary at a radial distance from Mars' centre of mass of 1,300–1,700 km.

Structure of the crust

Surface ages

Box 1 shows geological epochs on Mars, which, from oldest to youngest, are the Noachian, Hesperian and Amazonian, named after stratigraphic examples at type localities. Although it is possible to establish the relative timing of events during crustal evolution on the basis of stratigraphic superposition relationships⁴⁰ and globally distributed geophysical and geological observations⁴¹, there is a significant uncertainty as to when events occurred in an absolute time frame, because the initial impact flux is unknown. A variety of observations point to the fact that the predominance of planet-scale events occurred during the earliest martian epoch, the Noachian (see reviews in this issue by Jakosky and Phillips (pages 237–244), Stevenson (pages 214–219) and Baker (pages 228–236)), consistent with the concept of an 'accelerated' martian thermal evolution compared to Earth.

Crustal physiography

The first impression of the martian crust is that it exhibits a much different appearance in the northern and southern hemispheres (Fig. 2). This hemispheric dichotomy^{42,43} is manifest topographically as a contrast between the high-standing southern hemisphere and low-lying northern hemisphere, such that the south pole is nearly 6-km higher than the north pole^{44,45}. The dichotomy is also expressed in the surface geology, as the southern-hemisphere crust is heavily cratered and thus ancient (Noachian in age). In contrast, the northern hemisphere is smoother and displays far fewer craters, reflecting the fact that it has been resurfaced later, during the Hesperian epoch^{46,47}. The dashed line in Fig. 2 corresponds to the boundary between geologically older and younger provinces. The boundary zone is characterized by complex geology^{48,49} and offsets in regional elevation⁵⁰.

Another prominent global-scale feature is Tharsis, a vast, complex topographic rise that is a locus of volcanic activity and includes the

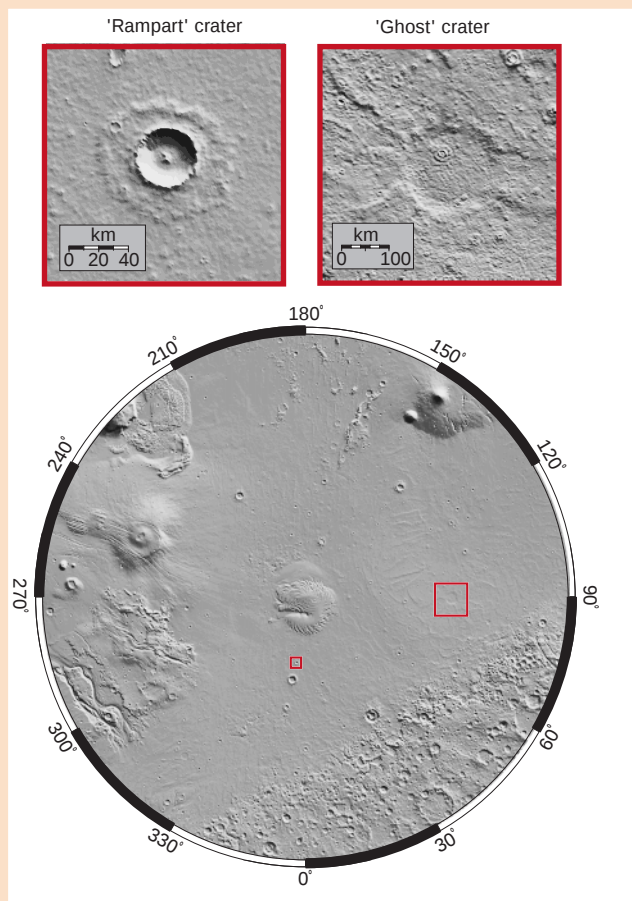


Figure 4 Shaded relief map centred on the north pole of Mars¹²¹. Insets show a 'rampart' crater whose unusual ejecta blanket represents evidence for the presence of subsurface volatiles at the time of the impact event, and a previously unrecognized, partially buried 'ghost' crater within the Utopia basin, which provides an estimate of the thickness of fill within Utopia since the buried crater formed.

largest martian shield volcanoes. The formation of Tharsis represented a principal source of stress in the martian crust and resulted in significant tectonism (fracturing and faulting), most prominently expressed by the massive Valles Marineris canyon system. The Elysium region in the northern lowlands is another crustal rise marked by volcanism and faulting, at a much smaller scale than Tharsis.

Mars also preserves the record of large impacts that occurred during the Noachian in the post-accretional phase of heavy bombardment. The largest preserved basins are Hellas, Utopia, Argyre and Isidis. The Utopia basin, buried beneath the plains of the northern hemisphere⁵¹, is about the same size as Hellas, but much shallower (2.5 km compared with 11 km)⁴⁵, which provides evidence for assessing the thickness of the northern-hemisphere fill in that region.

Crustal structure

As a result of the recent availability of globally distributed, high-resolution topography⁴⁵ and gravity⁵² data, it is now possible to map the subsurface structure of the martian crust and upper mantle⁵³. Inversions for planetary internal structure from gravity and topography are inherently non-unique and should, whenever possible, be developed in concert with additional constraints. The simplest plausible structure assumes a uniform-density crust and mantle, which provides a lower limit on the planet's mean crustal thickness and global crustal volume, but does not account for density variations within the crust or upper mantle. Figure 2 shows a crustal thickness model that was constructed under this assumption from the most up-to-date topography⁵⁴ and gravity⁵⁵ fields from MGS⁵⁶. The model

assumes a crustal density of $2,900 \text{ kg m}^{-3}$, consistent with plausible crustal compositions, and an assumed mantle density of $3,500 \text{ kg m}^{-3}$, consistent with bulk composition models³⁴. The mean thickness of the crust in Fig. 2 is $\sim 50 \text{ km}$, corresponding to 4.4% of the planetary volume.

Because of the absence of seismic measurements, the thickness of the martian crust cannot, as on the Earth and Moon, be 'anchored' by the depth of a seismic velocity discontinuity that characterizes the interface between layers of different compositions. Instead, the mean crustal thickness is constrained by calculations of the minimum value of viscosity of the lower crust that allows the observed relief at the crust–mantle boundary to be maintained over geological time^{53,57}. Different assumptions about the density difference between the crust and mantle or the crustal rheology or thermal structure could lead to variations in mean crustal thickness of as much as several tens of per cent. But for any reasonable assumptions, the mean thickness is inconsistent with earlier studies⁵⁸ that suggested a thickness of 100–250 km. The earlier estimates were based on geophysical inversions of poor-quality data⁵⁹ and the observation that martian meteorites exhibit a composition consistent with significant partial melting, which was interpreted as indicative of a large crustal volume^{58,60}.

An alternative model could invoke a crust of approximately uniform thickness that exhibits spatial variations in crustal and/or mantle density. Although possible, such a scenario is not easily reconciled with the observed progressive change in crustal composition from the southern hemisphere to the north. For example, the 'pole-to-pole' variation in crustal thickness shown in Fig. 2 could instead be explained by a 2% difference in the density of the mantle distributed over a depth range of 300 km. However, a planetary-scale density variation would be difficult to maintain over geological time against mantle flow, particularly early in martian history subsequent to the emplacement of the crust when mantle temperatures were likely to have been high. A warm interior would enhance crustal flow and cause variations in crustal thickness to relax away by viscous flow.

Figure 2 shows that, as for the Moon^{61,62}, the martian crust is thinned beneath all resolvable major impact basins, owing to a combination of excavation and crustal rebound during the impact process. The Tharsis province is characterized by a complex thick crust, representing massive ($\sim 3 \times 10^8 \text{ km}^3$) accumulation of magmatic and volcanic materials that point to volcanic construction as a significant contributor to its high-standing topography^{53,63}. The crust beneath Valles Marineris, which cuts a deep crustal exposure through eastern Tharsis, is thinned along the central axis, as in terrestrial rift zones. The interpretation of Tharsis as a massive volcanic pile is strengthened by the probable recognition in Valles Marineris that the full vertical extent of the canyon walls are composed of layered volcanic sequences (Fig. 3)⁶⁴. Both Elysium and the Solis Planum region in southern Tharsis appear as topographic rises with crustal roots that represent evidence of regionally enhanced melting.

Figure 2 also shows that Mars displays two crustal provinces that do not correlate entirely with the geological expression of the hemispheric dichotomy. One crustal province extends from the south polar region towards the north, and encompasses much of the southern highlands and the southern half of Tharsis. The second includes the northern lowlands and Arabia Terra. Although the boundary between crustal provinces correlates imperfectly with the boundary of the geological dichotomy, it does show some correlation with global-scale compositional units mapped by thermal emission observations⁶. On the basis of crustal structure, Arabia Terra may represent exposed basement of the resurfaced plains in the northern lowlands.

The northern lowlands

Before the launch of MGS, the northern lowlands had been viewed as vast, featureless plains, consisting of mainly Hesperian-aged volcanic flows^{46,47,65}. But high-resolution shaded-relief topography (Fig. 4)

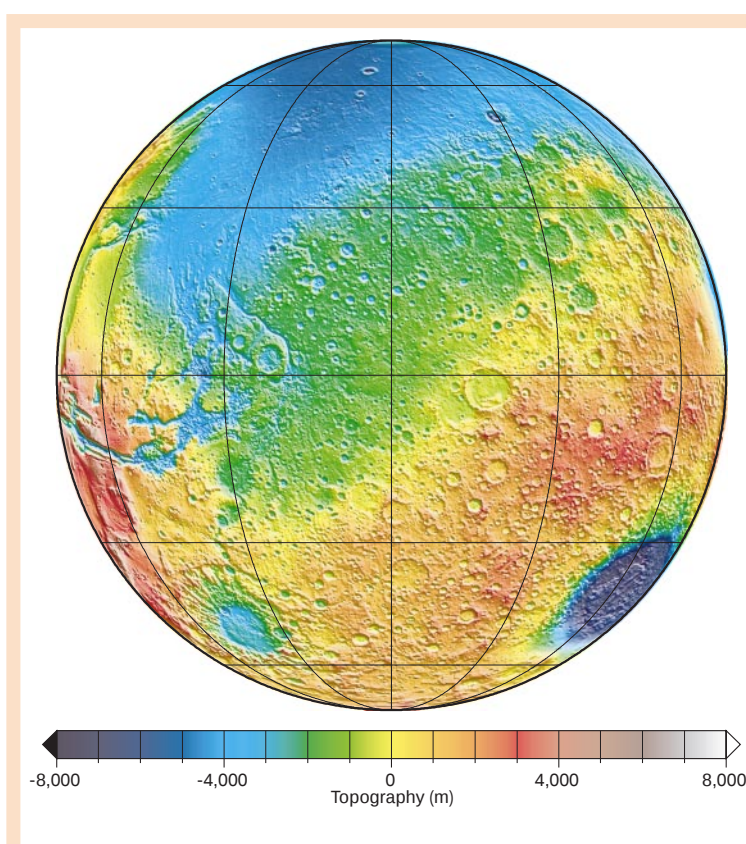


Figure 5 Spherical projection of Mars' topography⁵⁴ with 0° E longitude at front centre. The Hellas impact basin is located at the lower-right edge, with the low elevation of the basement depression shown in purple. The figure highlights the 2-km-high annulus of material that surrounds the basin and contributes to the elevation difference between Mars' southern and northern hemispheres.

now shows the lowlands to be rich in subtle detail, containing pervasive tectonic ridges^{66,67}, morphologic evidence for the flow of water far from the mouths of outflow channels⁵⁴, and 'ghost craters' that have been partially buried by plains material^{68,69}. Gravity and topography data now suggest a significant sedimentary contribution to the northern-hemisphere resurfacing⁵³.

The age of basement crust beneath the resurfaced northern plains and the timing of the resurfacing are key factors in understanding Mars' geological evolution. The buried Utopia basin demands that the crust beneath the northern plains is Noachian in age, as impactors of the required size would have been swept up by the planets shortly after accretion. Obscured and partially buried craters provide additional evidence for the age of the northern lowlands and the sequence of events in its history. The size–frequency distribution of topographic arcs resembling crater rims and large, shallow circular depressions⁶⁸ indicates that the number of craters beneath the northern plains is comparable to that in the southern highlands. Thus the northern lowlands beneath the later plains material are approximately as old as southern highlands — that is, Noachian in age.

Because crater diameter scales with depth⁷⁰, the relief of partially buried craters can be used to estimate the thickness of lowland fill. Recent analysis using this approach shows that the northern plains vary spatially in thickness⁶⁸. The density of partially buried craters within the much larger and deeper Utopia basin dictates that at least some of the resurfacing occurred not long after the formation of Utopia; if such craters had accumulated on the original floor of Utopia, the fill is sufficiently thick that they would not be detectable. It seems likely that much of the fill in Utopia consists of sediments transported from the southern highlands during the Noachian, when

water flowed freely on the surface (ref. 71, and see review in this issue by Leovy, pages 245–249). Massive erosion in the southern highlands⁷² may have been driven by volatiles associated with the formation of Tharsis in the Noachian⁷³. The deep fill in Utopia may contain material that eroded from the southern highlands, forming the dichotomy boundary scarp, which is developed prominently south of Utopia⁷⁴. These observations place important constraints on geological evolution models: any explanation for the northern lowlands must now explain both the low elevation and infilling, and the realization that these events occurred at different times.

Many impact craters on Mars have unusual ejecta blankets referred to as ‘ramparts’ (Fig. 4), which have been interpreted as evidence for impact into a water- or ice-rich substrate. Recent analysis of high-resolution Mars topography^{54,75} has shown that virtually all impact craters on the northern plains larger than about 3 km in diameter display rampart structures, which provides evidence that ice or water was ubiquitous in the subsurface of the northern hemisphere during the Hesperian epoch and possibly more recent times.

Mantle dynamics and thermal evolution

Lithosphere and heat flow

Models of the thermal evolution of Mars are sensitive to the distribution of internal heat sources and the style of mantle heat loss, which depends on internal structure and temperature distribution⁷⁶. Models of heat loss by thermal convection require assumptions about the concentrations of radioactive heat-producing elements (potassium, uranium and thorium) partitioned into the crust. These concentrations are controlled by the manner in which such large atoms partition into melt, as well as the fraction of the mantle that melted to generate the crust⁷⁷.

Joint analysis of gravity and topography has yielded estimates of the effective elastic thickness of the martian lithosphere (the outer shell of long-term strength). The elastic thickness is usually interpreted as the depth to an isotherm ($\sim 650^\circ\text{C}$) beneath which the martian interior is too weak to support stresses over geologically long ($\sim 10^8$ year) intervals. Measures of elastic thickness allow estimates of surface heat flow, which provide constraints on models of thermal evolution. Results reveal a significant variability in lithosphere structure among the main crustal provinces^{53,78}, but show a general trend: elastic thickness values generally decrease with increasing surface age, consistent with declining heat flux from the martian mantle with time. The southern highlands are oldest, indicative of crustal stabilization earliest in martian history, followed by the northern lowlands, which display an elastic thickness that reflects the thermal state during the period of northern-hemisphere resurfacing. Although Tharsis as a whole is old, it displays the thickest lithosphere, with implied heat flow reflecting the thermal state during the formation of the volcanoes, which formed relatively recently in comparison to the ancient and broad rise⁴⁶.

Thermal evolution

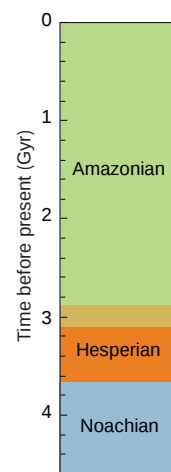
Thermal evolution models^{39,79,80}, constructed for a range of geometries, initial conditions, mantle viscosity structures and heat-production scenarios, all indicate a significantly declining heat flow with time. Uniform viscosity models³⁹ produce the slowest cooling during early martian evolution, whereas models with strongly temperature-dependent viscosity^{80–82} show the most rapid fall-off in early heat flow.

Sleep¹ has suggested that plate tectonics operated during the Noachian epoch and that the northern-hemisphere crust formed by crustal spreading. His model predicted a specific plate configuration that included active and passive margins, the latter of which correlate with the geological expression of the dichotomy. Although no obvious geophysical signature of the hypothesized long-extinct margins is preserved, it may be possible to test the hypothesis more specifically with recently acquired data. A northern-lowlands crust of thickness ~ 35 km (Fig. 2) could be produced by crustal spreading if

Box 1

Martian geological timescale

The martian geological record is inferred on the basis of mapped stratigraphy¹²³. Relative ages of stratigraphic units are based on superposition relationships. Absolute numerical ages can be assigned given models that relate the estimated impact-crater flux to the areal density of craters on distinct parts of the planetary surface. In discussing geological time, 1 Gyr is 10^9 years and 1 Myr is 10^6 years. The figure shows the martian stratigraphic epochs from oldest to youngest: the Noachian, Hesperian and Amazonian. Along with the stratigraphic record is a recent estimate of the absolute timescale¹²⁴.



the mantle temperature were somewhat greater than that typical of modern terrestrial mid-ocean ridges⁸³. However, it is unclear whether a such a thick crust could subduct and drive plate recycling, because the basalt–eclogite phase transition, which aids subduction, would not occur on Mars until a depth of ~ 200 km.

Formation of Tharsis

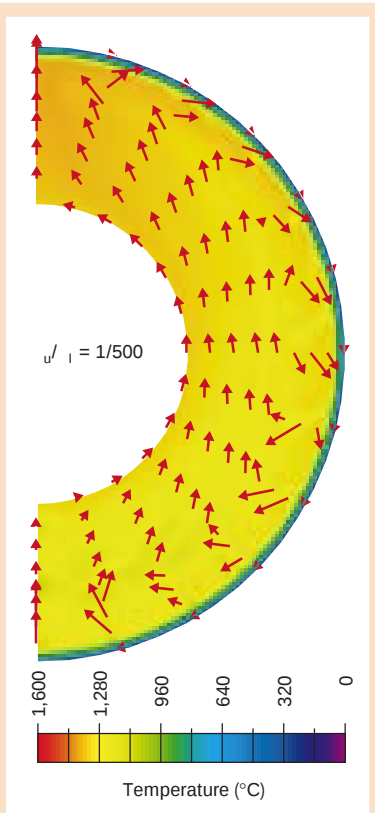
Various studies have used fault patterns and gravity/topography relationships^{84–88} in an attempt to understand the formation of Tharsis. As contributors to the formation of the province, data and models suggest a combination of volcanic construction and uplift, and heating associated with a major mantle plume or plumes. The presence or absence of a lower-mantle perovskite layer has significant implications for development of a hemispheric-scale plume. Convection in a martian mantle that contains the spinel-to-perovskite phase transition^{89–93} will suppress short-wavelength convective perturbations and preferentially develop long-wavelength flow. The velocity field may form a single upwelling plume that could potentially explain the formation of Tharsis, although models have not yet produced the desired pattern in sufficiently short timescales. A single thermal plume could provide the heat source required for extensive magmatism and volcanism, and could account for a fraction of the topographic uplift of Tharsis, although most of the current topography is probably due to piling up of volcanic material^{53,63,94}.

The location of Tharsis as antipodal to the Hellas impact basins has raised the suggestion that the impact could have initiated the mantle thermal anomaly that led to Tharsis. However, no model has demonstrated the plausibility of this idea. The location of Tharsis on the boundary of the northern and southern crustal-thickness provinces might not be coincidental. Models of continent–ocean boundaries on Earth⁹⁵ show a focusing of convective activity.

Formation of the crustal dichotomy

Hypotheses to explain the hemispheric dichotomy have included one or more massive impacts into the northern hemisphere^{96,97}, thinning of the northern-hemisphere crust by mantle convection^{98–100}, and an early period of tectonic-plate recycling¹. No assessments before MGS^{74,101} took into account the difference between the dichotomy as expressed in the surface geology and topography and that apparent in the crustal structure. Topography and crustal-structure data pose three challenges for impact-origin hypotheses, the first being the lack of global correlation of crustal thickness with the geological expression of the dichotomy (Fig. 2). Second, the topography (Fig. 4) and crustal thickness reveal a distinctly non-circular outline of the

Figure 6 Convection calculation illustrating concentration of heat loss in one hemisphere of Mars. The calculation uses a finite element approach and assumes axisymmetric geometry¹²². The figure shows temperature (in colour) and velocity (arrows) fields for a martian mantle with a 750-km-thick low-viscosity upper layer. The model is characterized by an 80-km-thick conductive lid, a viscosity contrast between the upper and lower mantle layers of 1/500, and a core size of half the planetary radius. Free slip and isothermal boundary conditions were imposed at the surface and crust–mantle boundary. The calculation includes internal heat production and temperature- and depth-dependent viscosity. The convection pattern results in preferential heating of one hemisphere that could potentially explain the dichotomy in crustal structure.



northern lowlands that cannot be explained easily by one or a few large impacts. And third, circular zones of crustal thinning that are characteristic of large impact basins are not apparent other than for Utopia. If the northern depression formed as a result of impact, it must have occurred very early in martian history (that is, before the formation of Utopia), and the topographic and crustal-thickness signatures must have been modified by subsequent processes.

A surprising result from MGS is that the geological expression of the dichotomy may be due in part to an impact in the southern hemisphere. Figure 5 illustrates the concentric nature of material around Hellas, which plausibly has been excavated during impact. The annulus stands 2 km above its surroundings and accounts for a significant amount of the high-standing topography of the southern hemisphere. Material excavated from Hellas represents a major re-distribution of the martian crust⁵³, contributing in part to the surficial expression of topography along part of the dichotomy boundary⁴⁵. However, Hellas' contribution to the geological dichotomy does not explain all the elevated topography in the southern hemisphere, nor does it account for the geological contrast between the hemispheres and the presence of a scarp at some parts of the boundary.

Mantle convection models for the origin of the dichotomy require a long-wavelength pattern of heat loss with upwelling in one hemisphere and downwelling in the other. Convective mechanisms have also been problematic because although hemispheric-scale mantle-flow models have been developed for the Moon¹⁰², it has been difficult to produce a long-wavelength pattern of heat loss in a planet with a core as large as Mars has^{39,103}. The reason for the difficulty has been the assumption that the mantle is of uniform viscosity. Hemispheric-scale convection is possible if, as for Earth, the martian mantle is layered in viscosity such that the upper mantle is at least 100 times less viscous than the lower mantle (Fig. 6). Such a layer could correspond to a martian asthenosphere, as occurs on Earth beneath the lithosphere. As the viscosity of the lower layer increases relative to the upper layer, deformation becomes more efficient at longer wavelengths. For conditions that may be plausible for early Mars, a

hemispheric pattern of heat loss develops in about 230 Myr and is maintained for over 1 Gyr, and could potentially explain early formation of crust in the Noachian as well as volcanic resurfacing of the northern lowlands in the Hesperian. A long-wavelength convective pattern cools the core and mantle more efficiently than shorter-wavelength convection and could have contributed to the demise of the core dynamo (ref. 104, and see review in this issue by Stevenson, pages 214–219).

Crustal magnetization

Vector magnetic-field measurements from MGS^{105–107} reveal broad, intense zones of alternating magnetic polarity in the southern-hemisphere crust (Terra Cimmeria) and several isolated anomalies in the northern lowlands (Fig. 1 in the review by Stevenson). The radial components of the anomalies have amplitudes an order of magnitude or more greater than anomalies on Earth measured at comparable altitude¹⁰⁸. On Earth, continental magnetic anomalies are thought to be mostly due to magnetic induction by the Earth's dipole field, whereas the oceanic crust displays magnetic stripes dominated by remanence. At present, Mars lacks an internally generated dipole magnetic field, and crustal anomalies most likely represent remanent magnetism from an internal field in the distant past^{105,109}. Stevenson (pages 214–219) presents evidence for a vigorous core dynamo that operated during the Noachian, before the end of heavy bombardment. Here we focus on the nature of crustal magnetization and its relationship to crustal formation.

The magnetic properties of crustal rocks are determined by their iron mineralogy¹¹⁰, which bears on crustal composition as well as formation or alteration conditions. The intensity of observed magnetizations on Mars implies that the magnetizing process is probably thermoremanence, which is more efficient at magnetizing rocks than remanence associated with crystallization or sediment deposition. The minerals responsible for intense magnetization of the martian crust are likely to be similar to the most magnetizing constituents in Earth's crust^{111,112}, with the martian examples characterized by higher average iron abundances. Favoured are minerals with high magnetic remanence, such as pyrrhotite (Fe_7S_8), magnetite (Fe_3O_4), titanomagnetite ($\text{Fe}_2\text{O}_3\text{-FeTiO}_3$), haematite (Fe_2O_3) and maghemite ($\gamma\text{-Fe}_2\text{O}_3$). Most of these minerals have been detected from MGS orbital spectroscopy¹¹³, lander surface analysis¹¹⁴ or martian meteorites^{115,116}. Their existence suggests a more oxidizing environment than that which characterizes the martian mantle³⁸, which suggests that aqueous alteration might have contributed to magnetization^{17,37}.

The timing of magnetization is related to the nature of crustal formation. The concentration of magnetization in southern-hemisphere crust and the absence of anomalies in the Hellas basin points to the early operation and subsequent demise of a dynamo early in martian history, implying rapid cooling of the crust (refs 105, 109, and the review by Stevenson). Magnetization of martian meteorite ALH84001 (ref. 116) is also consistent with an early dynamo. An alternative suggestion — that the dynamo turned on later in martian history¹¹⁷ — requires magmatic additions to, or thermal re-working of, the crust. Figure 2 shows that the crustal structure of the magnetized region in Terra Cimmeria does not seem to be distinctive from the non-magnetized parts of the southern hemisphere. There is no apparent evidence from the crustal structure for proposed intrusive magmatism, although such evidence is apparent in the Tharsis and Elysium regions. If there were later-stage intrusions or re-working of southern-hemisphere crust, the evidence must be at smaller spatial scales than the resolution of the crustal thickness model. The existence of localized magnetic anomalies in the northern lowlands and part of Tharsis, which display surface ages that range from Hesperian to Amazonian, has been invoked as additional evidence of a late turn-on of the martian dynamo¹¹⁷. But these anomalies do not correlate with surface topography and must be a consequence of buried sources^{52,53}, perhaps impacts, of likely Noachian age⁶⁸.

On the basis of amplitudes of magnetic anomalies and assuming a magnetic intensity comparable to that of Earth's upper oceanic crust¹¹⁰, the depth of the so-called Curie temperature for magnetization in Terra Cimmeria is the same as the thickness of the elastic lithosphere (~30 km)^{53,78}. Thus magnetics, gravity and topography provide similar estimates of the thermal state of the southern-hemisphere crust at the time of thermal stabilization.

Formation and evolution of the crust

Analysis of tungsten isotopes suggests that the martian core formed within the first 30 Myr of Solar System history¹¹⁸. The tungsten anomalies, which are controlled by metal-silicate fractionation and thus date core formation, correlate with neodymium anomalies, which are controlled by the formation of silicate magmas, and thus probably date mantle differentiation to form the crust. This correlation argues that core and crust formation were probably close to synchronous. These data also suggest that Mars has not experienced any significant large-scale convective mixing of the mantle or impact re-distribution of the crust since the time of core formation. However, it is difficult to reconcile this latter result with the evidence for a rapidly cooling (and therefore dynamically mixing) early mantle.

It has been suggested that, like the Earth's moon, primordial Mars was characterized by a 'magma ocean'¹¹⁹ that cooled to form the crust. But although limited possible evidence has been cited on the basis of isotopic signatures of the shergottites¹²⁰, surface geochemical evidence for such a phenomenon has not yet been identified. If Mars exhibited an early period of pervasive shallow melting, the thermal and compositional layering that would have resulted could have suppressed convective cooling of the mantle. It might also have allowed conductive cooling to significantly reduce temperatures at the base of the crust, allowing variations in crustal thickness to be preserved for a longer time.

Unsolved questions

Analyses of martian meteorites and the wealth of orbital data from MGS have greatly aided our understanding of the chemistry and structure of the martian crust and mantle, but many fundamental questions remain. What was the mechanism of crustal formation, and how is it related to planetary differentiation and the formation of the crustal dichotomy? Did the crust form with a structure similar to that currently observed, or did it develop more uniformly and evolve to the present configuration? How is the mantle stratified in mineralogy and viscosity? What is the temporal evolution of mantle dynamics and why did the planet cool so fast so early? Is there a causal relationship between Tharsis and the hemispheric dichotomy? What is the nature of highly magnetized crustal materials? Where are the carbonate and sulphate deposits that were expected to have formed in Mars' earlier, wetter environment? And what was the inventory and role of water in the geophysical, geological and climatological evolution of Mars? Progress on these questions can be attained by a well-balanced programme of theoretical and experimental investigations (see article in this issue by Carr and Garvin, pages 250–253), spacecraft exploration that includes orbiters, low-altitude platforms and *in situ* observations, and analysis of samples of the martian crust returned to Earth. □

1. Sleep, N. H. Martian plate tectonics. *J. Geophys. Res.* **99**, 5639–5655 (1994).
2. Bell, J. F. I. et al. Near-infrared imaging of Mars from HST: surface reflectance, photometric properties, and implications for MOLA data. *Icarus* **138**, 25–35 (1999).
3. Soderblom, L. A. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 557–593 (Univ. Arizona Press, Tucson, 1992).
4. Mustard, J. F. et al. The surface of Syrtis Major: composition of the volcanic substrate and mixing with altered dust and soil. *J. Geophys. Res.* **98**, 3387–3400 (1993).
5. Mustard, J. F. & Sunshine, J. M. Seeing through the dust: Martian crustal heterogeneity and links to the SNC meteorites. *Science* **267**, 1623–1626 (1995).
6. Bandfield, J. L., Hamilton, V. E. & Christensen, P. R. A global view of Martian surface composition from MGS-TES. *Science* **287**, 1626–1630 (2000).
7. McSweeney, H. Y. et al. Chemical, multispectral, and textural constraints on the composition and origin of rocks at the Mars Pathfinder site. *J. Geophys. Res.* **104**, 8679–8715 (1999).
8. McSweeney, H. Y. Jr SNC meteorites: clues to Martian petrologic evolution? *Rev. Geophys.* **23**,

- 391–416 (1985).
9. McSweeney, H. Y. Jr What have we learned about Mars from SNC meteorites? *Meteoritics* **29**, 757–779 (1994).
10. Sherman, D. M., Burns, R. G. & Burns, V. M. Spectral characteristics of the iron oxides with application to the Martian bright region mineralogy. *J. Geophys. Res.* **87**, 10169–10180 (1982).
11. Bell, J. F. in *Mineral Spectroscopy: A Tribute to Roger G. Burns* (eds Dyan, M. D., McCammon, C. & Schaefer, M. W.) (Geol. Soc. Spec. Pub. 5) 359–380 (Geol. Soc., Boulder, 1996).
12. Fanale, F. P., Postawko, S. E., Pollack, J. B., Carr, M. H. & Pepin, R. O. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 1135–1179 (Univ. Arizona Press, Tucson, 1992).
13. Morris, R. V. et al. Mineralogy, composition, and alteration of Mars Pathfinder rocks and soils: evidence from multispectral, elemental, and magnetic data on terrestrial analogue, SNC meteorite, and Pathfinder samples. *J. Geophys. Res.* **105**, 1757–1817 (2000).
14. Christensen, P. C., Bandfield, J. L., Smith, M. D., Hamilton, V. E. & Clark, R. N. Identification of a basaltic component of the Martian surface from Thermal Emission Spectrometer data. *J. Geophys. Res.* **105**, 9609–9621 (2000).
15. Christensen, P. R. et al. The Mars Global Surveyor Thermal Emission Spectrometer experiment: investigation description and surface science. *J. Geophys. Res.* (in the press).
16. Minititi, M. E. & Rutherford, M. J. Genesis of the Mars Pathfinder "sulfur-free" rock from SNC parental liquids. *Geochim. Cosmochim. Acta* **64**, 2535–2547 (2000).
17. McSweeney, H. Y. Jr et al. Geochemical evidence for magmatic water within Mars from pyroxenes in the Shergotty meteorite. *Nature* **409**, 487–490 (2001).
18. Noble, S. K. & Pieters, C. M. Type 2 terrain: compositional constraints on the Martian lowlands. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1230 <http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1230.pdf> (2001).
19. Jagoutz, E. Chronology of SNC meteorites. *Space Sci. Rev.* **56**, 13–22 (1991).
20. Chen, J. H. & Wasserburg, G. J. Formation ages and evolution of Shergotty and its parent planet from U-Th-Pb systematics. *Geochim. Cosmochim. Acta* **50**, 955–968 (1986).
21. Borg, L. E., Nyquist, L. E., Taylor, L. A., Weisman, H. & Shih, C.-Y. Constraints on Martian differentiation processes from Rb-Sr and Sm-Nd isotopic analyses of the basaltic shergottite QUE94201. *Geochim. Cosmochim. Acta* **61**, 4915–4931 (1997).
22. Bogard, D. D. & Johnson, P. Noble gases in the antarctic meteorite. *Science* **221**, 651–654 (1983).
23. Bogard, D. D., Nyquist, L. E. & Johnson, P. Noble gas contents of shergottites and implications for the Martian origin of SNC meteorites. *Geochim. Cosmochim. Acta* **48**, 1723–1739 (1984).
24. Swindle, T. D., Caffee, M. W. & Hohenberg, C. M. Xenon and other noble gases in shergottites. *Geochim. Cosmochim. Acta* **50**, 1001–1015 (1986).
25. Wiens, R. C., Becker, R. H. & Pepin, R. O. The case for a martian origin of the shergottites. II. Trapped and indigenous gas components in EETA79001 glass. *Earth Planet. Sci. Lett.* **77**, 149–158 (1986).
26. McKay, D. S. et al. Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. *Science* **273**, 924–930 (1996).
27. Mittlefehdt, D. W. ALH84001, a cumulate orthopyroxene member of the martian meteorite clan. *Meteoritics* **29**, 214–221 (1994).
28. Ash, R. D., Knott, S. F. & Turner, G. A 4-Gyr shock age for a martian meteorite and implications for the cratering history of Mars. *Nature* **380**, 57–59 (1996).
29. Treiman, A. H. A petrographic history of Martian meteorite ALH84001: two shocks and an ancient age. *Meteoritics* **30**, 294–302 (1995).
30. Treiman, A. H. The history of Allan Hills 84001 revised: multiple shock events. *Meteoritics Planet. Sci.* **33**, 753–764 (1998).
31. Wood, J. A. et al. in *Basaltic Volcanism on the Terrestrial Planets* (eds McGeechin, T. R., Pepin, R. O. & Phillips, R. J.) 634–699 (Pergamon, New York, 1981).
32. Bills, B. G. Geodetic constraints on the composition of Mars. *J. Geophys. Res.* **95**, 14131–14136 (1990).
33. Folkner, W. M., Yoder, C. F., Yuan, D. N., Standish, E. M. & Preston, R. A. Interior structure and seasonal mass redistribution of Mars from radio tracking of Mars Pathfinder. *Science* **278**, 1749–1752 (1997).
34. Dreibus, G. & Wanke, H. Mars: a volatile-rich planet. *Meteoritics* **20**, 367–382 (1985).
35. Longhi, J., Knittle, E., Holloway, J. R. & Wanke, H. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 184–208 (Univ. Arizona Press, Tucson, 1992).
36. Bertka, C. M. & Fei, Y. Implications of Mars Pathfinder data for the accretion history of the terrestrial planets. *Science* **281**, 1838–1840 (1998).
37. Wadhwa, M. Redox state of Mars' upper mantle and crust from Eu anomalies in shergottite pyroxenes. *Science* **291**, 1527–1530 (2001).
38. Wanke, H. & Dreibus, G. Chemical composition and accretion history of the terrestrial planets. *Phil. Trans. R. Soc. Lond. A* **325**, 545–557 (1988).
39. Schubert, G. & Spohn, T. Thermal history of Mars and the sulfur content of its core. *J. Geophys. Res.* **95**, 14095–14104 (1990).
40. Tanaka, K. L., Scott, D. H. & Greeley, R. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 345–382 (Univ. Arizona Press, Tucson, 1992).
41. Solomon, S. C. et al. What happened when on Mars? Some insights into the timing of major events from Mars Global Surveyor data. *EOS Trans. Am. Geophys. Union* P31A-06 (2001).
42. Mutch, T. A., Arvidson, R. E., Head, J. W., Jones, K. L. & Saunders, R. S. *The Geology of Mars* (Princeton Univ. Press, Princeton, 1976).
43. Carr, M. H. *The Surface of Mars* (Yale Univ. Press, New Haven, 1981).
44. Smith, D. E. & Zuber, M. T. The shape of Mars and the topographic signature of the hemispheric dichotomy. *Science* **271**, 184–188 (1996).
45. Smith, D. E. et al. The global topography of Mars and implications for surface evolution. *Science* **284**, 1495–1503 (1999).
46. Scott, D. H. & Tanaka, K. L. Geologic map of the western equatorial region of Mars. US Geol. Survey Map I-1802-A (1986).
47. Greeley, R. & Guest, J. E. Geologic map of the eastern equatorial region of Mars. US Geol. Survey Map I-1802-B (1987).
48. Sharp, R. P., Soderblom, L. A., Murray, B. C. & Cutts, J. A. The surface of Mars 2. Un cratered terrains. *J. Geophys. Res.* **76**, 331–342 (1971).
49. Sharp, R. P. Mars: fretted and chaotic terrains. *J. Geophys. Res.* **78**, 4073–4083 (1973).
50. Frey, H., Sakimoto, S. E. & Roark, J. The MOLA topographic signature at the crustal dichotomy boundary. *Geophys. Res. Lett.* **25**, 4409–4412 (1998).

51. McGill, G. E. Buried topography of Utopia, Mars: persistence of a giant impact depression. *J. Geophys. Res.* **94**, 2753–2759 (1989).
52. Smith, D. E. *et al.* The gravity field of Mars: results from Mars Global Surveyor. *Science* **286**, 94–97 (1999).
53. Zuber, M. T. *et al.* Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* **287**, 1788–1793 (2000).
54. Smith, D. E. *et al.* Mars Orbiter Laser Altimeter (MOLA): experiment summary after the first year of global mapping of Mars. *J. Geophys. Res.* (in the press).
55. Lemoine, F. G. *et al.* An improved solution of the gravity field of Mars (GMM-2B) from Mars Global Surveyor. *J. Geophys. Res.* (in the press).
56. Albee, A. A., Palluconi, F. D. & Arvidson, R. E. Mars Global Surveyor mission: overview and status. *Science* **279**, 1671–1672 (1998).
57. Nimmo, F. & Stevenson, D. Estimates of Martian crustal thickness from viscous relaxation of topography. *J. Geophys. Res.* **106**, 5085–5098 (2001).
58. Sohl, F. & Spohn, T. The interior structure of Mars: implications from SNC meteorites. *J. Geophys. Res.* **102**, 1613–1635 (1997).
59. Janle, P. Bouguer gravity profiles across the highland-lowland escarpment on Mars. *Earth, Moon Planets* **28**, 55–67 (1983).
60. Spohn, T., Sohl, F. & Breuer, D. Mars. *Astron. Astrophys. Rev.* **8**, 181–236 (1998).
61. Bratt, S. R., Solomon, S. C., Head, J. W. & Thurber, C. H. The deep structure of lunar basins: implications for basin formation and modification. *J. Geophys. Res.* **90**, 3049–3064 (1985).
62. Zuber, M. T., Smith, D. E., Lemoine, F. G. & Neumann, G. A. The shape and internal structure of the Moon from the Clementine mission. *Science* **266**, 1839–1843 (1994).
63. Solomon, S. C. & Head, J. W. Evolution of the Tharsis province of Mars: the importance of heterogeneous lithospheric thickness and volcanic construction. *J. Geophys. Res.* **82**, 9755–9774 (1982).
64. McEwen, A. S., Malin, M. C., Carr, M. H. & Hartmann, W. K. Voluminous volcanism on early Mars revealed in Valles Marineris. *Nature* **397**, 584–586 (1999).
65. Greeley, R. & Spudis, P. D. Volcanism on Mars. *Rev. Geophys.* **19**, 13–41 (1981).
66. Head, J. W., Kreslavsky, M. A. & Pratt, S. Northern lowlands of Mars: evidence for widespread volcanic flooding and tectonic deformation in the Hesperian period. *J. Geophys. Res.* (in the press).
67. Withers, P. & Neumann, G. A. Enigmatic northern plains of Mars. *Nature* **410**, 651 (2001).
68. Frey, H., Shockey, K. M., Frey, E. L., Roark, J. H. & Sakimoto, S. E. H. A very large population of likely buried impact basins in the northern lowlands of Mars revealed by MOLA data. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1680 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1680.pdf>> (2001).
69. Kreslavsky, M. A. & Head, J. W. III Stealth craters in the northern lowlands of Mars: evidence for a buried Early-Hesperian-aged unit. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1001 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1001.pdf>> (2001).
70. Pike, R. J. Depth/diameter relations of fresh lunar craters: revision from spacecraft data. *Geophys. Res. Lett.* **1**, 291–294 (1974).
71. Squyres, S. W. & Kasting, J. F. Early Mars: how warm and how wet? *Science* **265**, 744–749 (1994).
72. Hynek, B. M. & Phillips, R. J. Evidence for extensive denudation of the Martian highlands. *Geology* **29**, 407–410 (2001).
73. Phillips, R. J. *et al.* Ancient geodynamics and global-scale hydrology on Mars. *Science* **291**, 2587–2591 (2001).
74. McGill, G. E. & Dimitriou, A. M. Origin of the Martian global dichotomy by crustal thinning in the late Noachian or early Hesperian. *J. Geophys. Res.* **95**, 12595–12605 (1990).
75. Garvin, J. B. & Frawley, J. J. Geometric properties of Martian impact craters: preliminary results from the Mars Orbiter Laser Altimeter. *Geophys. Res. Lett.* **25**, 4405–4408 (1998).
76. Sleep, N. H. Evolution of the mode of convection within terrestrial planets. *J. Geophys. Res.* **105**, 17563–17578 (2000).
77. Parmentier, E. M. & Zuber, M. T. Relaxation of crustal thickness variations on Mars: implications for thermal evolution. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1357 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1357.pdf>> (2001).
78. McGovern, P. J. *et al.* Gravity/topography admittances and lithospheric evolution of Mars: the importance of finite-amplitude topography. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1804 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1804.pdf>> (2001).
79. Schubert, G., Solomon, S. C., Turcotte, D. L., Drake, M. J. & Sleep, N. H. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 147–183 (Univ. Arizona Press, Tucson, 1992).
80. Grasset, O. & Parmentier, E. M. Thermal convection in a volumetrically heated, infinite Prandtl number fluid with strongly temperature-dependent viscosity: implications for planetary thermal evolution. *J. Geophys. Res.* **103**, 18171–18181 (1998).
81. Choblet, G., Grasset, O., Parmentier, E. M. & Sotin, C. Mars thermal evolution revisited. *Lunar Planet. Sci. Conf. XXX*, Abstr. 1556 <<http://www.lpi.usra.edu/meetings/LPSC99/pdf/1556.pdf>> (1999).
82. Spohn, T. Thermal histories of the terrestrial planets revisited. *EOS Trans. Am. Geophys. Union* **80**, P32B-05 (1999).
83. Forsyth, D. W. Crustal thickness and the average depth and degree of melting in fractional melting models of passive flow beneath mid-ocean ridges. *J. Geophys. Res.* **98**, 16073–16079 (1993).
84. Banerdt, W. B., Phillips, R. J., Sleep, N. H. & Saunders, R. S. Thick shell tectonics on one-plate planets: applications to Mars. *J. Geophys. Res.* **87**, 9723–9733 (1982).
85. Banerdt, W. B., Golombek, M. P. & Tanaka, K. L. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 249–297 (Univ. Arizona Press, Tucson, 1992).
86. Bills, B. G. & Nerem, R. S. A harmonic analysis of Martian topography. *J. Geophys. Res.* **100**, 26314–26326 (1995).
87. Harder, H. Mantle convection and the dynamic geoid of Mars. *Geophys. Res. Lett.* **27**, 301–304 (2000).
88. Anderson, R. C. & Golombek, M. P. Timing of Tharsis uplift from the distribution of tectonic structures. *J. Geophys. Res.* (in the press).
89. Weinstein, S. A. The effects of a deep mantle endothermic phase transition. *J. Geophys. Res.* **100**, 11719–11728 (1995).
90. Harder, H. & Christensen, U. A one-plume model of Martian mantle convection. *Nature* **380**, 507–509 (1996).
91. Breuer, D., Zhou, H., Yuen, D. A. & Spohn, T. Phase transitions in the Martian mantle: implications for the planet's volcanic evolution. *J. Geophys. Res.* **101**, 7531–7542 (1996).
92. Breuer, D., Yuen, D. A. & Spohn, T. Phase transitions in the Martian mantle: implications for partially layered convection. *Earth Planet. Sci. Lett.* **148**, 457–469 (1997).
93. Breuer, D., Yuen, D. A., Spohn, T. & Zhang, S. Three-dimensional models of Martian mantle convection with phase transitions. *Geophys. Res. Lett.* **25**, 229–232 (1998).
94. Zhong, S. Long-wavelength topography and gravity anomalies of Mars and their implications to the formation of the Tharsis rise. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 2124 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/2124.pdf>> (2001).
95. King, S. D. & Ritsema, J. African hot spot volcanism: small-scale convection in the upper mantle beneath cratons. *Science* **290**, 1137–1140 (2000).
96. Wilhelms, D. E. & Squyres, S. W. The Martian hemispheric dichotomy may be due to a giant impact. *Nature* **309**, 138–140 (1984).
97. Frey, H. V. & Schultz, R. A. Large impact basins and the mega-impact origin for the crustal dichotomy on Mars. *Geophys. Res. Lett.* **15**, 229–232 (1988).
98. Lingenfelter, R. E. & Schubert, G. Evidence for convection in planetary interiors from first-order topography. *Moon* **7**, 172–180 (1973).
99. Wise, D. U., Golombek, M. P. & McGill, G. E. Tectonic evolution of Mars. *J. Geophys. Res.* **84**, 7934–7939 (1979).
100. Wise, D. U., Golombek, M. P. & McGill, G. E. Tharsis province of Mars: geologic sequence, geometry, and a deformation mechanism. *Icarus* **35**, 456–472 (1979).
101. McGill, G. E. & Squyres, S. W. Origin of the Martian crustal dichotomy: evaluating hypotheses. *Icarus* **93**, 386–393 (1991).
102. Zhong, S., Parmentier, E. M. & Zuber, M. T. On the dynamic origin and asymmetric distribution of mare basalts. *Earth Planet. Sci. Lett.* **177**, 131–141 (2000).
103. Schubert, G., Bercovici, D. & Glatzmaier, G. A. Mantle dynamics in Mars and Venus: influence of an immobile lithosphere on three-dimensional mantle convection. *J. Geophys. Res.* **94**, 14105–14129 (1990).
104. Nimmo, F. & Stevenson, D. J. Influence of early plate tectonics on the thermal evolution and magnetic field of Mars. *J. Geophys. Res.* **105**, 11969–11979 (2000).
105. Acuna, M. H. *et al.* Global distribution of crustal magnetization discovered by the Mars Global Surveyor MAG/ER experiment. *Science* **284**, 790–793 (1999).
106. Connerney, J. E. P. *et al.* Magnetic lineations in the ancient crust of Mars. *Science* **284**, 794–798 (1999).
107. Purucker, M. *et al.* An altitude-normalized magnetic map of Mars and its interpretation. *Geophys. Res. Lett.* **27**, 2449–2452 (2000).
108. Langel, R. A., Phillips, J. D. & Horner, R. J. Initial vector magnetic anomaly map from MAGSAT. *Geophys. Res. Lett.* **9**, 269–272 (1982).
109. Acuna, M. H. *et al.* Results of the Mars Global Surveyor magnetic field investigation. *J. Geophys. Res.* (in the press).
110. Dunlop, D. J. & Ozdemir, O. *Rock Magnetism* (Cambridge Univ. Press, Cambridge, 1997).
111. Kletetschka, G., Wasilewski, P. J. & Taylor, P. T. Unique thermoremanent magnetization of multidomain sized hematite: implications for magnetic anomalies. *Earth Planet. Sci. Lett.* **176**, 469–479 (2000).
112. Kletetschka, G., Wasilewski, P. J. & Taylor, P. T. Mineralogy of the sources for magnetic anomalies on Mars. *Meteoritics Planet. Sci.* **35**, 895–899 (2000).
113. Christensen, P. C. *et al.* Detection of crystalline hematite mineralization on Mars by the Thermal Emission Spectrometer: evidence for near-surface water. *J. Geophys. Res.* **105**, 9623–9642 (2000).
114. Hargraves, R. B. *et al.* Magnetic enhancement on the surface of Mars? *J. Geophys. Res.* **105**, 1819–1827 (2000).
115. Cisowski, S. M. Magnetic studies on Shergotty and other SNC meteorites. *Geochim. Cosmochim. Acta* **50**, 1043–1048 (1986).
116. Weiss, B. P., Vali, H., Baudenbacher, F. J., Stewart, S. T. & Kirschvink, J. L. Records of an ancient Martian magnetic field. *Nature* (submitted).
117. Schubert, G., Russell, C. T. & Moore, W. B. Timing of the Martian dynamo. *Nature* **408**, 666–667 (2000).
118. Lee, D.-C. & Halliday, A. N. Core formation on Mars and differentiated asteroids. *Nature* **388**, 854–857 (1997).
119. Hess, P. C. & Parmentier, E. M. Implications of magma ocean cumulate overturn for Mars. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1319 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1319.pdf>> (2001).
120. Blichert-Toft, J., Gleason, J. D., Telouk, P. & Albareda, F. The Lu-Hf isotope geochemistry of shergottites and the evolution of the Martian crust-mantle systems. *Earth Planet. Sci. Lett.* **173**, 25–37 (1999).
121. Neumann, G. A., Rowlands, D. S., Lemoine, F. G., Smith, D. E. & Zuber, M. T. The crossover analysis of MOLA altimetric data. *J. Geophys. Res.* (in the press).
122. Zhong, S. & Zuber, M. T. Degree-1 mantle convection and the crustal dichotomy on Mars. *Earth Planet. Sci. Lett.* **189**, 75–84 (2001).
123. Tanaka, K. L. The stratigraphy of Mars. *J. Geophys. Res.* **91**, E139–E158 (1986).
124. Hartmann, W. K. & Neukum, G. Cratering chronology and the evolution of Mars. *Space Sci. Rev.* (in the press).

Acknowledgements

I thank G. Neumann and O. Aharonson for assistance with the crustal thickness inversion, H. McSweeney and W. Hartmann for providing material used in the figures, and J. Head, H. McSweeney and N. Sleep for helpful reviews. I also acknowledge helpful discussions with M. Parmentier, T. Grove, B. Weiss and members of the Mars Global Surveyor Laser Altimeter and Radio Science teams. This work was supported by the Mars Global Surveyor Project of the NASA Mars Exploration Program.

Water and the martian landscape

Victor R. Baker

Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona 85721-0011, USA (e-mail: baker@hwr.arizona.edu)

Over the past 30 years, the water-generated landforms and landscapes of Mars have been revealed in increasing detail by a succession of spacecraft missions. Recent data from the Mars Global Surveyor mission confirm the view that brief episodes of water-related activity, including glaciation, punctuated the geological history of Mars. The most recent of these episodes seems to have occurred within the past 10 million years. These new results are anomalous in regard to the prevailing view that the martian surface has been continuously extremely cold and dry, much as it is today, for the past 3.9 billion years. Interpretations of the new data are controversial, but explaining the anomalies in a consistent manner leads to potentially fruitful hypotheses for understanding the evolution of Mars in relation to Earth.

When the Mariner 9 spacecraft went into orbit on 14 November 1971, the surface of Mars was shrouded in a global dust storm. Fortunately, by March of 1972, the atmosphere cleared and the true complexity of the Mars landscape was finally revealed to the spacecraft's vidicon cameras. In addition to the immense volcanoes of Tharsis, there was a great equatorial canyon system, Valles Marineris, named for the discovery spacecraft. Most remarkable, however, were sinuous channels and valleys, from whose morphology it was inferred that water had once flowed on the surface of this now dry, unearthly cold planet. The late Harold Masursky, science team leader for the vidicon imaging experiment, wrote in 1973 (ref. 1):

The possible fluvial channels may record episodes when water was much more abundant in the atmosphere than it is at present. Planet-wide warmer interglacial periods would release not only the water locked in the polar caps but also that frozen in the subsurface as permafrost. Similar warmer and colder periods also are characteristic of terrestrial history.

Although intended primarily as support to landers seeking evidence of martian life, the Viking orbiters of the late 1970s returned 52,603 images of Mars, most of them at much higher resolution than the 7,329 images returned by Mariner 9. The pictures with the highest resolution have a pixel spacing of 7.5 m, although most frames resolve to several tens of metres. Viking orbiter images (Fig. 1) provided the basis for an understanding of water and landscape that prevailed until the past few years².

In 1997, the Mars Global Surveyor (MGS) spacecraft was inserted into Mars' orbit, but a variety of problems prevented it from achieving a circular mapping orbit until 19 February 1999. Two instruments are particularly relevant to the scientific study of the Mars landscape. The Mars Orbiter Camera (MOC) achieves a resolution of 1.4 m per pixel, but the required high data volume limits scenes to a kilometre or so in width at the highest resolutions³. The Mars Orbiter Laser Altimeter (MOLA) maps the topography of Mars with a precision better than 10 m (ref. 4). Together these instruments provide new data for studying Mars' landforms (Fig. 2), although human reasoning about those landforms remains a matter of long-standing scientific experience.

The study of Earth-like planetary surfaces — geomorphology — is not a disjointed collection of observational facts solely with which to test, or against which to constrain, theoretical models. Rather, such scientific inquiry proceeds from the informed colligation of landform observations to

the discovery of consistency and coherence, and, ultimately, to consilience⁵ in the theoretical accounting (explanation) of those observations. The key element of this inquiry is the formulation of one or more working hypotheses⁶, which are most often suggested (but not proved) by analogies of form and context among landscapes of known origin and those under scrutiny⁷. In the retroductive inferences of geomorphology^{8,9}, analogy serves merely to suggest fruitful working hypotheses, thereby leading to completely new theories that bind together any newly discovered facts. Mars' landscape provides particularly stimulating opportunities to practise geomorphological reasoning, generating hypotheses that may initially strike some researchers as outrageous¹⁰. Nevertheless, it is the productive pursuit of such hypotheses that leads ultimately to new understanding, not only of Mars, but also of Earth itself.

The surface of Mars is today extremely cold and dry. The atmosphere at the land surface is over 100 times less dense than that of Earth, and it holds only minuscule amounts of water vapour. For the present obliquity (tilt of the planet's rotational axis with respect to the orbital plane) of 25°, the residual north polar ice cap sublimates water in the northern spring/summer, and the vapour moves to and condenses at south polar areas, a pattern that may reverse in northern autumn/winter¹¹. Given its minor role in comparison to that on Earth, it is not surprising that on present-day Mars, water is replaced by wind as the most continuously active surface-modifying process¹².

In stark contrast to the current environment, however, numerous landforms provide signs, or indicators, of extensive past activity of water and ice on the martian surface. From the densities of impact craters on the various terrains it is possible to work out a chronology for this activity, such that Mars is divided stratigraphically into ancient heavily cratered uplands that formed prior to about 3.5 billion years ago (the Noachian epoch), intermediate cratered plains (the Hesperian epoch), and more lightly cratered areas (the Amazonian epoch) (see review in this issue by Zuber, pages 220–227). Water and ice were active on the surface during all these periods. The mode, timing and long-term cycling of water in surficial processes are the phenomena to be considered from the various signs of activity.

Signs of subsurface water and ice

Landforms indicative of ground ice in permafrost (perennially frozen ground) on Mars have been known since the early flyby missions of the 1960s. Images from Viking orbiters provided an overwhelming list of permafrost and ground-



Figure 1 Glaciated terrain east of Hellas Planitia, at latitude 42° S, longitude 252° W. This image from a Viking orbiter mission shows a scene about 180 × 140 km². At the top is the lineated valley fill of Reull Vallis, which may be an extant or relict debris-covered glacier about 10 km wide. The uplands at the centre and bottom of the image were eroded to produce forms typical of glacial alpine sculpture⁹⁴. At the base of several uplands are prominent lobate debris aprons. The longest of these, near the centre of the image, extends for 40 or 50 km from the brightly illuminated walls of the sculpted uplands. Prominent flow lineations show that the debris moved by viscous flow, probably facilitated by the plastic deformation of underlying ice. The lack of craters on the flow-lineated surfaces indicates a remarkably recent (possibly continuing) occurrence of the responsible flow processes.

ice indicators, including various kinds of patterned ground, thermokarst, hillslope features and mass-movement phenomena^{13,14}. Although originally attributed to permafrost processes, the immense polygons, 3–20 km across, of the northern Mars plains are much larger than the contraction-crack polygons typical of terrestrial permafrost terrains. These features are now explained variously as the tectonic uplift of basin floors, perhaps following removal of load from an overlying standing body of water¹⁵, or as the result of Rayleigh convection driven by unstable density or temperature gradients in a catastrophic flood deposit positioned over frozen ground¹⁶. Recently acquired very high-resolution MOC imagery reveals extensive areas of the northern plains and southern highlands of Mars, notably on crater floors (Fig. 3), where small-scale contraction-crack polygons, tens or hundreds of meters across, cover the landscape¹⁷. This polygonal terrain, which closely mimics the ice-wedge polygonal terrains of terrestrial permafrost areas, is essentially uncratered, indicating a surprisingly youthful phase of water-related activity on Mars.

Viking pictures revealed that many martian craters have a unique morphology, different from that observed elsewhere in the Solar System. Ejecta surrounding these craters are layered, and each layer has an outer edge terminating in a low ridge or escarpment. Named 'rampart craters', the flow-ejecta morphology most likely represents the incorporation of groundwater and ground ice¹⁸, although atmospheric effects on ejecta emplacement may also be important¹⁹. Thus, layered ejecta morphologies²⁰ can be used to characterize the past presence of water in various martian terrains²¹.

Figure 2 High-resolution Mars Orbiter Camera (MOC) image of a fluvial channel system at latitude 7.9° N, longitude 205.8° W, south of Cerberus Rupes (MOC Image M21-01914). The scene shows an area about 4 km across. A complex of anastomosing channels and streamlined uplands reveals a history of differential fluid erosion of layered bedrock and progressive degradation that produced terrace levels and abandoned spillways. Regularly spaced (about 60-m wavelength) rib-like bedforms are developed transverse to the direction of fluid flow in some of the channels. All these features are best explained by large-scale water flow. The lack of impact craters on the flood-scoured surfaces indicates that this flow occurred very recently in martian geological history. (Image provided courtesy of Malin Space Science Systems.)



First revealed by Mariner 9, the long and complex volcanic history of Mars contains a wealth of examples of interactions among volcanism, ice and water. Much as in the Pleistocene landscapes of Iceland, martian volcanism has produced features interpreted as table mountains built up on products of sub-ice eruptions, outburst flood channels, and extensive pyroclastic landscapes²². Some of the youngest volcanic landscapes occur in the Cerberus Rupes and Marte Vallis region²³, where cataclysmic flood channels²⁴ and volcanic lava flows²⁵ occur in close spatial and temporal association. Even Olympus Mons, one of the largest known volcano constructs, has a morphology interpreted by some to be indicative of water/ice volcanic interactions²⁶. Its huge aureole deposits, extending 1,000 km, may represent immense submarine landslides²⁷, similar to those of Earth's Hawaiian Islands²⁸.

Signs of surface water

The heavily cratered martian highlands are locally dissected by integrated networks of tributaries with widths of about 10 km or less, and lengths from <5 km to nearly 1,000 km. Drainage densities are generally much lower than for terrestrial valley networks²⁹, although an interesting set of valleys on martian volcanoes is more similar to terrestrial valleys in their general morphology and degree of terrain dissection³⁰. Among the other important morphological attributes of martian valley networks are the following: theatre-like valley heads, prominent structural control, low junction angles, quasi-parallel patterns, hanging valleys, irregular widening and narrowing,

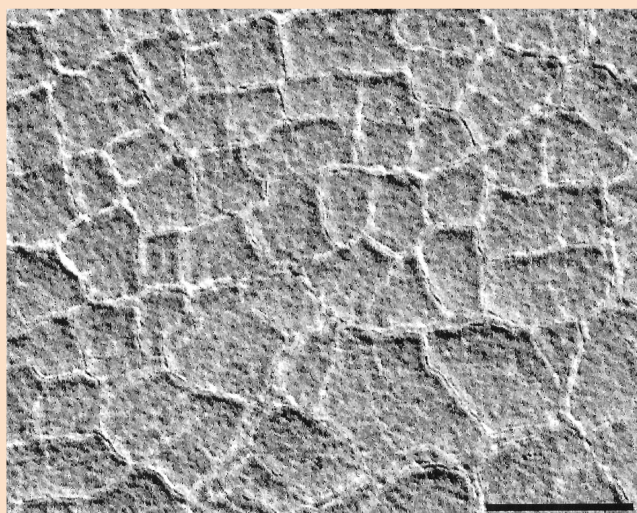


Figure 3 Large contraction-crack polygons developed on the floor of a northern-plains crater at approximately 67.5° N, 312.5° W (MOC Image M01-00294). Note the parallel bright linear ridges or troughs that comprise the boundaries of the polygons. This morphology is typical for terrestrial ice-wedge polygons that develop by the repeated seasonal or episodic melting, freezing and movement of water in the active layer overlying an ice-rich permafrost zone⁸⁸. Scale bar, 200 m.

and indistinct terminal areas³¹. In general, this assemblage of features seems best explained by groundwater sapping processes³², excellent examples of which occur in the sandstone terrains of the Colorado plateau³³ and in parts of the Hawaiian Islands³⁴.

Although most of the networks occur in the heavily cratered terrains of Mars³⁵, as many as 25–35% of the valleys may be Hesperian or Amazonian in age³⁶. Moreover, MOC images show that some valleys, which formed early in Mars' history, were later reactivated by smaller flows — a discovery that had been anticipated from Viking data³⁷. Another observation, the localized development and incomplete dissection of upland areas by valleys³⁸, is reinforced by the lack of fine-scale tributaries to valleys observed at MOC resolution³⁹. Explanations for such relationships remain contentious, with lack of precipitation as only one of several alternatives, others of which include the probable high infiltration capacities of the martian surfaces³⁷, the burial of small-scale features beneath mantling deposits, and the local hydrothermal⁴⁰ and/or snowmelt processes³².

Many valley networks have orientations consistent with the effects of surface deformation by extensive volcanic loading at Tharsis, inferred to have occurred by late Noachian time⁴¹. These networks occur in heavily cratered terrains that show evidence of extensive erosion during the Noachian^{42,43}. The average rates for this presumed water erosion are estimated at 10^2 – 10^4 mm per 10^6 years, which are comparable to the lower range of terrestrial values. But post-Noachian rates are estimated at only 10^{-1} – 10^{-2} mm per 10^6 years (ref. 44). This discrepancy led some researchers to postulate an epoch of late Noachian climate conducive to water erosion, followed by post-Noachian conditions that precluded such erosion^{2,45}. However, if post-Noachian aqueous erosion episodes are limited to very short episodes⁴⁶, then the average post-Noachian erosion will yield the same rate as calculated.

The term 'channel' is properly restricted to one class of martian trough-like landforms that display at least some evidence for large-scale fluid flow on their floors. The principal landform of interest here is the outflow channel, which shows evidence of flows emanating from zones of regional collapse known as 'chaotic terrain'. The martian outflow channels are immense (Fig. 4), as much as 150 km wide and 2,000 km in length. It was recognized shortly after their discovery that they possessed a suite of bedforms and morphological

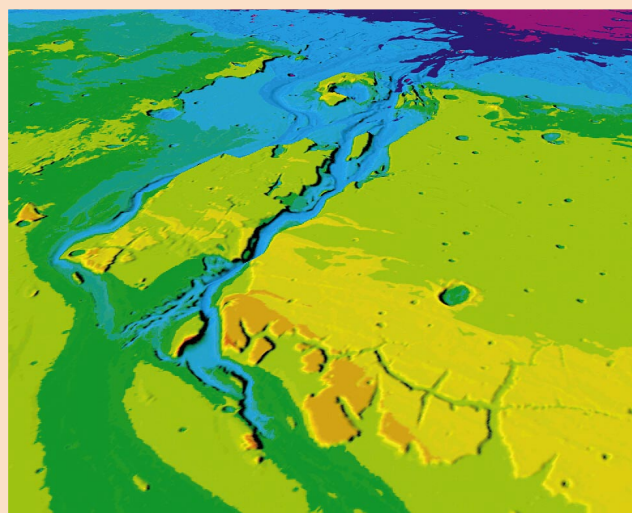


Figure 4 Oblique view of topographic data from the Mars Orbiter Laser Altimeter (MOLA) showing downstream portions of the outflow channel Kasei Vallis. The channelled area extends about 2,000 km from Echus Chasma, which lies to the south (bottom) of the image shown here. Flows in Kasei Vallis extended over a width of about 150 km (bottom of the image) before they turned right (east), as shown at the left centre of the scene. Note the deeply incised sinuous channels and prominent streamlining and scouring of residual uplands (top centre of the image). The northern (left) channel of this incised zone is about 50 km wide, and it is also illustrated by the simplified cross-section at the top of Fig. 5. Elevation is indicated by the colours, progressing from dark blue at –4,000 m, to dark green at about –1,000 m, to dark orange at +2,000 m. (Image produced by T. M. Hare, US Geological Survey, Flagstaff, Arizona.)

relationships similar to what is exhibited in the 'Channeled Scabland'⁴⁷, a terrestrial landscape created by Pleistocene glacial-related cataclysmic flooding in the northwestern coterminous United States⁴⁸.

At their largest scale, the outflow channels are broadly anastomosing and split by residual uplands, or 'islands', of pre-flood-modified terrain (Fig. 2). The channels have low sinuosity and high width–depth ratios. Pronounced flow expansions and constrictions occur, as do prominent divide crossings, hanging valleys and structural control of erosion. At a finer scale, streamlining of the residual uplands is very well developed, as are longitudinal grooves, inner channels, cataract complexes, scabland and bar complexes. Although a variety of other fluid-flow systems have been invoked to explain these features, the whole assemblage of these landforms is best explained by cataclysmic flood processes, with particular analogy to the origin of the Channeled Scabland^{31,49}. Nevertheless, it is also recognized that some important differences derive from the peculiar martian environment, notably its lower gravitational acceleration, its much lower atmospheric pressure and its prevailing subfreezing temperature in comparison to Earth. These special conditions on Mars would influence such factors as sediment transport mechanics⁵⁰, cavitation and ice formation⁵¹, debris flow⁵² and possible large-scale ice processes in the channels⁵³.

The sizes of martian outflow channels imply immense discharges of water, exceeding any known flood flows on Earth (Fig. 5). The calculated peak-flow magnitudes are comparable to those of the high-discharge western boundary currents of Earth's world ocean, such as the Gulf Stream, the Kuroshio and the Agulhas⁵⁴. These currents are integral to Earth's climate system, distributing heat poleward from equatorial latitudes. Major disruption of Earth's circulation pattern provides an endogenetic model for large-scale terrestrial climate change⁵⁵. By analogy, Mars' climate could have been impacted by the megaflood discharges transferring water and heat from the equatorial Tharsis volcanic province to the northern plains of the planet⁴⁶.

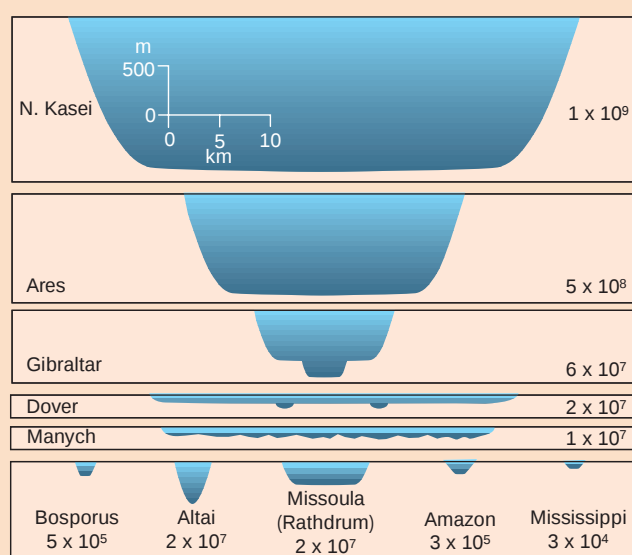


Figure 5 Comparison of simplified channel cross-sections for cataclysmic flood channels on Mars (upper two channels) and Earth (lower channels). Peak discharges are indicated in cubic metres per second. Gibraltar, Dover and Bosporus are straits that were inundated by flows that spilled between marine basins. Manych, Altai and Missoula (Rathdrum) are associated with glacial meltwater floods that developed at the end of the last ice age. The two modern rivers, Amazon and Mississippi, are dwarfed by the ancient flood channels. More detail on the floods associated with these channels is provided by Baker⁵⁴.

Outflow activity was probably initiated in the Noachian⁵⁶, and outflow events occurred during the early Hesperian⁵⁷, the late Hesperian/early Amazonian^{58,59} and, surprisingly, during the past 10 million years²³. The latter floods occurred in the Marte Vallis/Cerberus Rupes region⁶⁰ and the Tharsis area⁶¹. Thus, principal fluvial events (and possibly associated climate change) occurred right up to most recent martian history. Alternatively, however, outflow channel events are considered to be random outbursts from a progressively thickening cryosphere⁶².

Integral to the controversy over post-Noachian water-related climate change on Mars is the lack of understanding of how such immense discharges can be released from the martian subsurface. The role of the cryosphere seems important, perhaps for generating pressurized aquifer confinement⁶³. Dike emplacement and penetration of the ice-rich cryosphere⁶⁴, pressurized volcanic influences on fracture systems⁶⁵ and decompression of gas hydrates deep in the crust^{66,67} are other possible factors in the cataclysmic release of outburst floods.

The largest outflow channels delivered their immense discharges to the northern plains of Mars. Evidence of temporary inundation of the northern plains was recognized on the basis of Viking data^{40,68–70}, and some aspects of the inundation hypothesis are consistent with new results from the MOLA instrument⁷¹. These include the extreme topographic smoothness of the northern plains, the apparent burial of wrinkle ridges, correlation of basin topography with landforms of likely aqueous origin, gradation of outflow-channel floors to ancient inundation levels, consistency of inundation volumes with possible Mars water balances, and some correspondence of one of the possible shoreline levels to a deformed ancient equipotential surface⁷¹. However, landforms interpreted as diagnostic of shoreline processes⁷² are not confirmed by detailed study of MOC imagery⁷³. Named 'Oceanus Borealis'⁴⁶, the northern-plains inundation remains controversial as to its size, age (or ages), persistence and episodic formation. Other interpretations of the northern-plains inundation include the emplacement of massive debris flows⁷⁴, effects of glaciation⁷⁵ and a complex layering of both ice- and water-laid deposits⁷⁶.



Figure 6 MOC image (M09-04718) of small gullies and other hillslope features in the central peak area of Hale Crater (latitude 36° S, longitude 37° W). Scale bar, 200 m.

Evidence for past lakes abounds in numerous impact craters on Mars⁷⁷. Most of the lakes formed in the Upper Hesperian and Lower Amazonian⁷⁸. Lacustrine terraces and Gilbert-type deltas indicate that the lakes must have persisted for 10^3 to 10^4 years (ref. 79). Standing water in lakes for such timescales requires a major change in climate from present-day conditions, although ice-covered lakes could persist for considerable time at current conditions or for those associated with the range of variation in Mars' orbital parameters⁸⁰. The relationships at the Gale Crater palaeolake are particularly interesting because of extensive work^{81,82} indicating a relatively young age. However, the first interpretation of MOC images claimed that the Gale Crater deposits and all martian lake deposits were Noachian in age⁸³. This ancient age designation was even applied to extensive sedimentary deposits in the Valles Marineris, an interpretation that has been strongly challenged following public release of the images^{84,85}. Ironically, the claim of a Noachian age for the deposits requires their extensive exhumation late in martian history in order to account for the lack of impact craters on the exposed surfaces. Such a presumed erosion event would require even more extreme martian climate change than would the aqueous emplacement of the deposits that many investigators originally associated with post-Noachian lacustrine activity^{77,79–81}. This important scientific controversy continues.

Perhaps the most striking discoveries made from MOC images are the very young, relatively small-scale debris-flow gullies (Fig. 6), initially explained as the result of groundwater seepage and subsequent surface runoff⁸⁶. The gullies have morphologies essentially identical to those that are common on terrestrial hillslopes in periglacial environments⁸⁷, such as coastal Greenland, Iceland, Svalbad and arctic

Canada. On Earth, initiation of debris flow occurs when soil water saturation follows surface melting of snow cover or ground ice⁸⁸. Remarkably, the Mars gullies generated debris-flow deposits that are superimposed on aeolian bedforms and on small-scale contraction-crack polygons⁸⁶, all of which are uncratered. Thus, the various age indicators show the gullies and debris flows to be extremely young, probably active within the past several million years of Mars' history.

The profound implications of a recent Earth-like origin for the young gullies led some researchers to seek alternative mechanisms that might generate debris flows on Mars. Ideally, the modelled processes should be able to occur in the current cold, dry environment. For example, the existence of CO₂ vapour-supported debris flows might, in theory, permit build-up of a liquid-CO₂ aquifer behind a near-surface dry-ice barrier that is subsequently breached by heating⁸⁹. In one extreme scenario, Mars is hypothesized as both currently being and always having been an extreme cryogenic world in which water behaves largely as a solid mineral, and nearly all manifestations of fluid-flow phenomena involve CO₂ gas or liquid⁹⁰. This 'white Mars' model postulates a history of mean temperatures continuously lower than those prevailing today.

Signs of glacial ice

Given that periglacial landscapes on Earth involve warmer climatic conditions than do glacial landscapes⁹¹ (Fig. 7), it is interesting that glacial interpretations of martian landforms have generally been more controversial than have periglacial interpretations. A strong case was made for the contribution of glacial processes to the origin of landforms associated with certain outflow channels⁹², and both glacial and flood outflow processes can operate together or sequentially in large-scale channel morphogenesis⁹³. Nevertheless, the scale of the presumed climate changes occurring late in Mars' history that would have been necessary to account for glaciation led many researchers to express extreme scepticism in regard to any glacial interpretation².

The most intricate assemblages of glacial landforms are hypothesized to explain landscapes in the mountainous uplands adjacent to the Argyre and Hellas impact basins⁹⁴. Local alpine glacial sculpture in the Charitum Montes, south of Argyre at latitude 57° S, includes cirques, horns, aretes, grooves and V-shaped valley troughs. Many of the cirques and troughs seem to be occupied by extant or relict rock glaciers or debris-covered glaciers (see discussion below). The adjacent floor of Argyre Planitia has landforms that can be interpreted as moraines, drumlins, esker-like ridges, kettles, outwash plains and glaciolacustrine plains⁹⁴.

The esker-like ridges (Fig. 8), which are 10–200 km long, have proven to be a source of lively controversy at professional meetings. Taken as individual landforms, these sinuous ridges have been variously explained as sand dunes, inverted stream valley fills, igneous or clastic dikes, lake beach ridges, mudflow levees and the wrinkle ridges generated by deep-seated faults. MOC images of the Argyre Planitia sinuous ridges (Fig. 9) reveal apparent sedimentary strata, boulders and discontinuous ridges, with heights of 10–100 m, and widths of 200–2,000 m. Together with the associated glacial landforms, the esker interpretation seems to provide a consistent hypothesis explaining all relevant observations.

Even more prominent esker-like ridges are associated with the Hesperian-aged Dorsa Argentea Formation at latitudes 75–80° S. Numerous morphological similarities show these ridges to be similar to terrestrial eskers⁹⁵. The eskers seem to have been produced by the meltback of ice-rich deposits, perhaps deriving from an extensive south polar ice cap in the middle (Hesperian) portion of Mars' history⁹⁶. The drainage from this ice cap was carried through prominent valleys to the Argyre impact basin, where it constituted a temporary lake.

MOC images of lineated valley fills in the northern fretted terrain and valleys draining to Hellas, plus images of the lobate debris aprons in uplands near Hellas (Fig. 1) and Argyre, show features interpreted to represent very recent glacial flow. These include crevasse-like

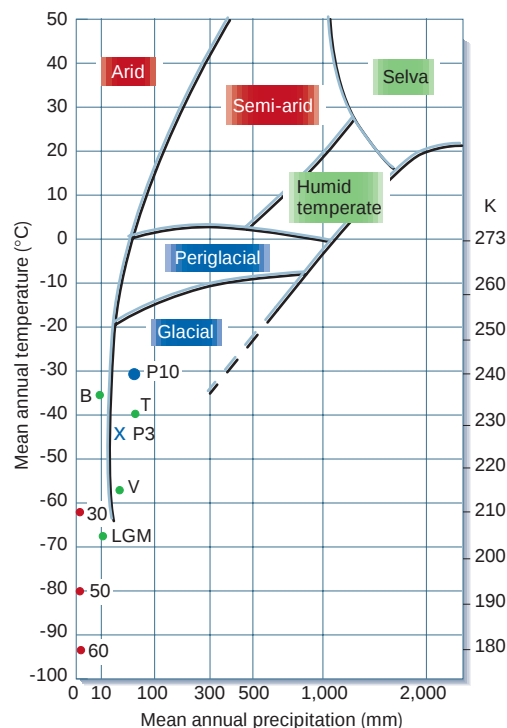


Figure 7 Morphogenetic regions for climate-related landforms on Earth⁹¹. Note the expanded precipitation scale at lower values. Modern terrestrial glacial environments include Taylor Dome (T) and Vostok (V) in Antarctica. Beacon Valley (B) occurs in the cold-arid, nonglaciated Dry Valley region of Antarctica. Conditions inferred from the last glacial maximum (LGM) in Antarctica were most similar to those of modern middle latitudes on Mars (30, 50 and 60 are the Mars conditions at the indicated latitudes). Modelled conditions¹¹⁸ for ancient Mars at 300 mbar (300 hPa) atmospheric pressure (P3) and 1,000 mbar (P10) could have been comparable to the cold glacial conditions on modern Earth.

fracture concentrations and medial moraines displaying glacier-like patterns of tributary convergence and downvalley flow. The surfaces of the lineated valley fills and lobate debris aprons are uncratered, indicating likely emplacement within the past several million years. MOLA profiles of lobate debris aprons in Deuteronilus Mensae (latitude 40° N) and Protonilus Mensae (latitude 46° N) reveal shapes best explained by solid-state deformation of ice hundreds of metres thick⁹⁷. However, at current Mars surface temperatures and very low accumulation rates (<1 cm yr⁻¹), flow rates for these large ice masses would be so slow⁹⁸ that they could not be produced on the timescale of 10⁶–10⁷ years implied by their uncratered surfaces. Because the hypothesized glaciers are associated with outwash and other water-drainage landforms⁹⁴, higher past temperatures are implied, and these could achieve Earth-like strain rates in the deforming ice.

One school of glacial geomorphological thought holds that a continuum exists on Earth from true glaciers to debris-covered glaciers, and then to remobilized talus or till⁹⁹. The last phenomena would certainly be classed as rock glaciers, which another geomorphological school holds strictly to be permafrost phenomena¹⁰⁰. The prevailing hypothesis for the Mars lineated valley fills and lobate debris aprons follows the permafrost theory in its claim that deformation is achieved in ground ice that cements slope-derived debris¹⁰¹. In contrast, following the continuum theory, one can hypothesize that recently active, true martian glaciers carved the V-shaped troughs, cirques, aretes and other landforms of local highland terrains. These ancient glaciers would have had large accumulation areas, requiring atmospheric transport of water to the site and a net surplus of input from snow to offset losses by sublimation and meltwater runoff.

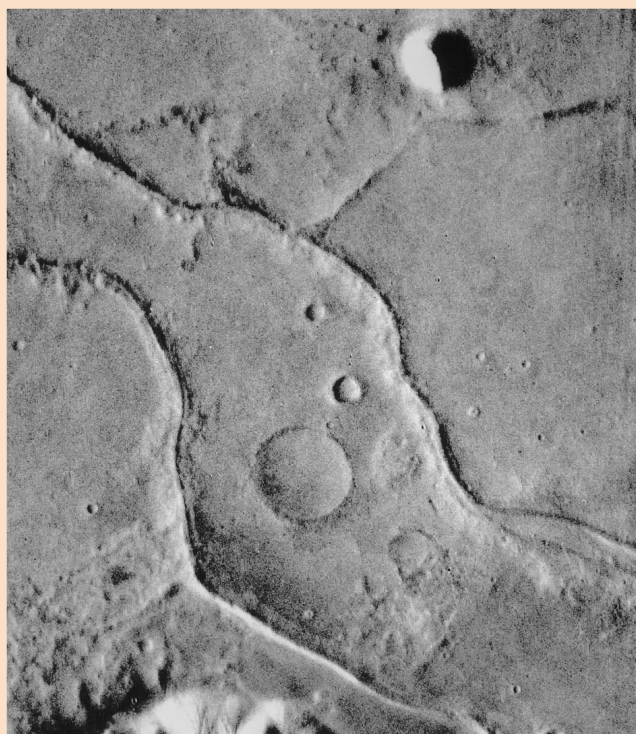


Figure 8 Esker-like ridges in southern Argyre Planitia (latitude 56° S, longitude 40° W). This Viking orbiter image (567B33) shows a scene about 50 km across. The ridges are sharp-crested and about 1,000 m in width.

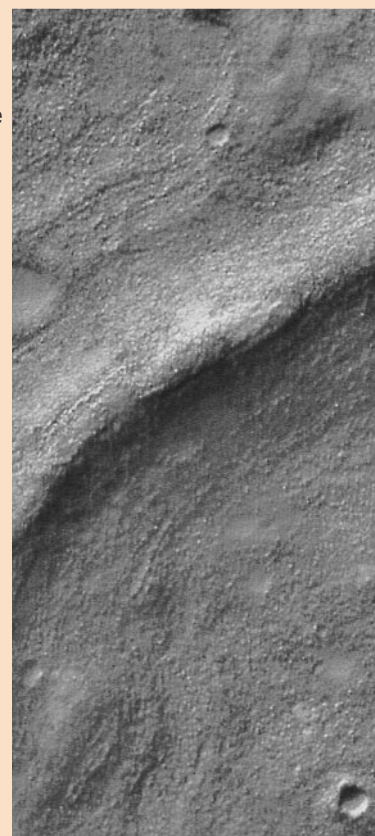
When climate shifted to more adverse conditions, such that input of snow no longer exceeded losses, the glaciers wasted back. But debris accumulating on the now extensive ablation zone would have greatly reduced sublimation loss. Moreover, if temperatures also fell, meltwater loss would cease. The resulting debris-covered glacier might still have had a very small accumulation zone at its head, and this could sustain continued flow of ice at very slow rates, as observed in rock glaciers on Earth¹⁰².

For the extremely cold conditions that currently prevail on Mars, debris-covered ice would have sublimation rates as low as 10^{-5} cm yr⁻¹ (ref. 103). Thus, even very small accumulation rates over small areas might sustain a positive net mass balance, although flow rates would be extremely slow in a glacier at a mean annual temperature of -75 °C (the current condition on Mars at about 45° latitude). Of course, present water precipitation at this latitude is effectively zero, and water condensation occurs today primarily at near-polar latitudes.

The current mean annual temperature at Vostock on the East Antarctic Ice Sheet is -57 °C, and the accumulation rate is only 2 cm yr⁻¹. During the Last Glacial Maximum, ice-core data indicate that mean annual temperatures may have been as much as 20 °C colder, and accumulation rates may have been less than 1 cm yr⁻¹ (ref. 104). These conditions are very close to those on Mars today (Fig. 7), although Earth atmospheric pressures are, of course, much higher. In Beacon Valley, part of the Dry Valley region of Antarctica, precipitation is currently less than 1 cm yr⁻¹ water equivalent, and the mean annual temperature is -35 °C. The area is ice-free, but relict glacial ice occurs beneath a 0.5-m debris cover that is dated at 8.1 million years (Myr) old¹⁰⁵. The indicated sublimation rate is 10^{-5} cm yr⁻¹, the same rate noted above for debris-covered ice on Mars. With present-day conditions on Mars so close to those in parts of Antarctica (Fig. 7), it is reasonable to invoke a climate-change scenario to explain the observed glacial phenomena.

Temperatures achieved during the Last Glacial Maximum are representative only of Earth's most recent glacial epoch, the Late Cenozoic. Earlier periods of prolonged, extensive glaciations

Figure 9 Portion of MOC Image MOO-01511 showing detail of esker-like ridge in southern Argyre Planitia (Fig. 8). This image shows a scene about 3 km in width.



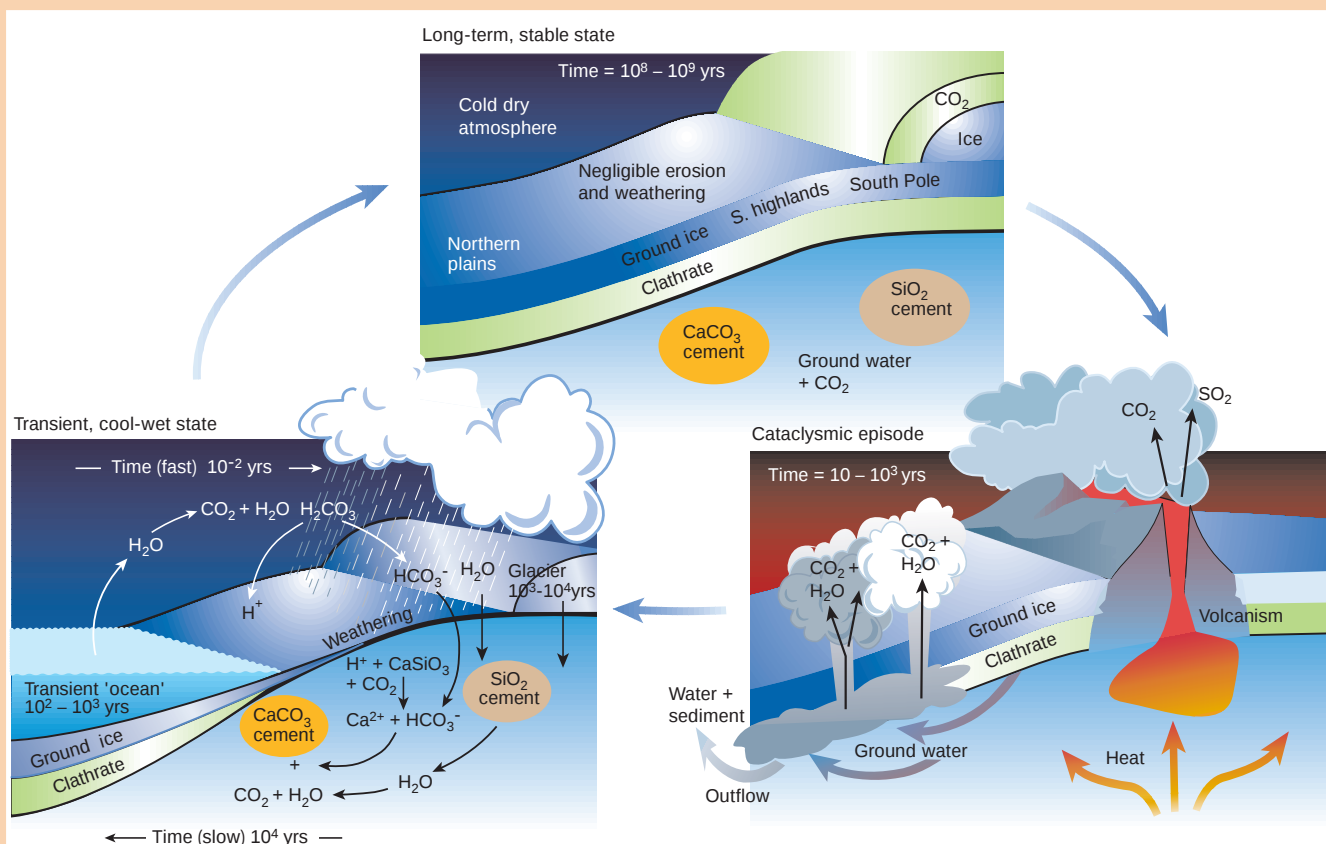
(megaglaciations) of Earth occurred in the Late Palaeozoic (~260–340 Myr ago), Late Devonian/Early Carboniferous (~350–360 Myr ago), Ordovician/Silurian (~430–450 Myr ago), Late Proterozoic (~500–900 Myr ago), Early Proterozoic (about 2,200–2,400 Myr ago) and Archaean (~2,910–2,990 Myr ago)¹⁰⁶. The Late Proterozoic glaciation is particularly enigmatic, as there is evidence that Earth may have switched into Mars-like icehouse conditions by freezing the surface of the global ocean¹⁰⁷. The extreme glaciation is explained alternatively as a result of a huge change in planetary obliquity¹⁰⁸ or as a runaway cooling that terminates cataclysmically with a super-greenhouse build-up caused by volcanism¹⁰⁹. The rapid melting of ice and weathering of the land surface lead to carbonate precipitation in the ocean, thereby terminating the super-greenhouse.

The glacial epochs of Earth are explained by two classes of theoretical models: external and internal. External causes include the orbital parameters, axial tilt (obliquity), precession and eccentricity, all of which acted as the 'pacemaker' of Pleistocene glaciations¹¹⁰. Internal causes include the role of conveyor-belt oceanic circulation⁵⁵ and the arrangement of continental relief on the planet¹¹¹. For Mars, it is possible to imagine an internal scenario capable of explaining the diverse observations summarized in this paper (see Box 1), although there have been many objections raised to an earlier version of this model². Alternatively, the complex variability of Mars' obliquity over long timescales^{112,113} may be able to mobilize water¹¹⁴ in ways that will explain relatively recent manifestations of glacial action, lakes, fluvial gullies and the melting of near-surface ground ice.

Is it possible, as a theoretical matter of pure physics, that the numerous martian landscapes, temporal associations of landforms, and proximal-to-distal relationships — all analogous to water-generated process-form relationships on Earth — could have been produced by a fluid that mimics the behaviour of water? For appropriate relationships with temperature and pressure, CO₂ is a possible candidate for achieving the necessary physical behaviour⁹⁰. Moreover, the CO₂ hypothesis is parsimonious in the sense of being

Box 1

Cyclic and episodic change on Mars



A genetic model, presented in an early form in 1991 (ref. 46), predicts both long-term (10^8 – 10^9 years) and short-term (10^3 – 10^4 years) aspects of a global martian hydrological cycle (see figure above). The episodic inundation of the northern plains of Mars is one component of the short-term portion of the cycle. The hypothesis explains the phenomenally long epochs of post-heavy-bombardment time during which the Mars surface clearly had extremely cold and dry conditions similar to those prevailing today. The alteration of this prevailing condition with quasi-stable, short-duration warmer (cool) and wetter conditions explains many young landforms that require such conditions, including glaciated terrain, crater lakes, melting of permafrost, and other water-related landforms that occur throughout Mars' history.

Internal planetary heat could have provided the trigger for the massive outflows that transformed martian climate during the geologically short epochs of climatic change. Superimposed on the long-term monotonic decline in mantle heat flux for Mars, one can predict short episodes of higher heat flux to the surface. These episodes of higher heat flow seem consistent with the magmatic and tectonic history of Mars¹¹⁹.

The huge floods would have acted to initiate climate change through release of CO_2 , leading to greenhouse warming¹²⁰. During the short-duration thermal episodes of cataclysmic outflow, a temporary cool-wet climate would prevail. Water that evaporated off a transient 'ocean' would be transferred to uplands, including the Tharsis

volcanoes and portions of the southern highlands, where precipitation as snow would promote the growth of glaciers. However, such a cool-wet climate is inherently unstable for Mars. Water from the evaporating ocean would be lost to storage in highland glaciers, and through infiltration into the porous lithologies of the martian surface. This would lead to the demise of warm conditions owing to the progressive loss of CO_2 through dissolved gas in infiltrating acidic water, and through silicate weathering carrying bicarbonate into the subsurface by infiltration. Subsequent underground carbonate deposition would then release CO_2 to the groundwater, which would become trapped beneath an ice-cemented permafrost zone. The latter develops as the greenhouse effect declines, because of atmospheric loss of water and CO_2 over a timescale of 10^3 – 10^5 years. Concurrent decline in planetary heat flow, following the triggering peak episode, would lead to a downward extending permafrost zone that would progressively incorporate the recharging water and groundwater. As the permafrost extended downward into the stability field for CO_2 clathrate, this gas hydrate would accumulate above the gas-charged groundwater. Thus, the underground sequestering of clathrate, gas-charged groundwater and carbonate cements comprise the long-term reservoir for carbon on Mars. Only occasionally, and for relatively short duration, does carbon get transferred to the atmosphere, as greenhouse-promoting CO_2 , during cataclysmic ocean-forming episodes triggered by pulses of high heat flow. The whole process would be cyclic as illustrated in the diagrams.

capable of generating flow phenomena for the current near-surface environment of Mars. If Mars, unlike Earth, were not perturbed over geological time by major epochs of climate change, then a perpetual 'white Mars' might well occur. Nevertheless, it would be remarkable that this white Mars landscape evolved locally to generate surprisingly detailed copies of what on Earth is readily produced by the long-term

action of water and ice. It is also a measure of simplicity in a hypothesis that the most natural explanation, in this case an aqueous origin, tends to accord with greater consistency to the observed phenomena.

In choosing to explore the consequences of a water-based explanation for Mars' geomorphology, and the associated implications for both ancient and very recent climate change, one is not assured of

easily testable hypotheses. The extensive hydrosphere implied by past aqueous activity on Mars may only be extant as ground ice in the thick permafrost zone and as underlying groundwater. Yet, this is the type of environment in which the extremophile progenitors of Earth's biosphere probably evolved¹¹⁵. Indeed, early Mars provided an arguably better habitat for the inception and incubation of early life than did early Earth¹¹⁶. Episodic, brief episodes of aqueous activity on the martian surface may have exposed this biosphere to produce possible fossil indicators of its existence. More speculatively, a deep subsurface biosphere containing methanogenic archaea could have produced methane that accumulated beneath a growing martian cryosphere. The result could destabilize the martian cryosphere (see Box 1) and perhaps change the climate on short timescales¹¹⁷.

Ultimately in geomorphological investigation, one chooses the working hypothesis that is most fruitful with regard to connection to other lines of scientific inquiry, while still providing consistency and coherence in explaining the whole complex of available observations. The water-generated landforms and landscapes of Mars have proven especially difficult to unite under a single working hypothesis. Nevertheless, whatever explanation one chooses as a working hypothesis, its roots in the complex detail of Mars' landscape features ensure that it will provide a highly probable basis for further productive inquiry. □

1. Masursky, H. An overview of geologic results from Mariner 9. *J. Geophys. Res.* **78**, 4037–4047 (1973).
2. Carr, M. H. *Water on Mars* (Oxford Univ. Press, New York, 1996).
3. Malin, M. C. *et al.* Early views of the Martian surface from the Mars Orbiter Camera of Mars Global Surveyor. *Science* **279**, 1681–1685 (1998).
4. Smith, D. E. *et al.* The global topography of Mars and implications for surface evolution. *Science* **284**, 1495–1502 (1999).
5. Whewell, W. *Philosophy of the Inductive Sciences* (Parker, London, 1847).
6. Chamberlin, T. C. The methods of Earth-sciences. *Pop. Sci. Mon.* **66**, 66–75 (1904).
7. Gilbert, G. K. The origin of hypotheses, illustrated by a discussion of a topographic problem. *Science* **3**, 1–13 (1896).
8. Baker, V. R. in *The Scientific Nature of Geomorphology* (eds Rhoads, B. L. & Thorn, C. E.) 57–85 (Wiley, New York, 1996).
9. Baker, V. R. The pragmatic roots of American Quaternary geology and geomorphology. *Geomorphology* **16**, 197–215 (1996).
10. Davis, W. M. The value of outrageous geological hypotheses. *Science* **63**, 463–468 (1926).
11. Jakosky, B. M. & Haberle, R. M. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 969–1016 (Univ. Arizona Press, Tucson, 1992).
12. Greeley, R., Lancaster, N., Lee, S. & Thomas, P. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 730–766 (Univ. Arizona Press, Tucson, 1992).
13. Carr, M. H. & Schaber, G. G. Martian permafrost features. *J. Geophys. Res.* **82**, 4039–4054 (1977).
14. Lucchitta, B. K. Mars and Earth: comparison of cold climate features. *Icarus* **45**, 264–303 (1981).
15. Hiesinger, H. & Head, J. W. III Characteristics and origin of polygonal terrain in southern Utopia Planitia, Mars: results from Mars Orbiter Laser Altimeter and Mars Orbiter Camera data. *J. Geophys. Res.* **105**, 11999–12022 (2000).
16. Lane, M. D. & Christensen, P. R. Convection in a catastrophic flood deposit as the mechanism for the giant polygons on Mars. *J. Geophys. Res.* **105**, 17617–17627 (2000).
17. Seibert, N. M. & Kargel, J. S. Small-scale Martian polygons: liquid surface water. *Geophys. Res. Lett.* **28**, 899–902 (2001).
18. Strom, R. G., Croft, S. K. & Barlow, N. G. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 383–423 (Univ. Arizona Press, Tucson, 1992).
19. Schultz, P. H. & Gault, D. E. Atmospheric effects on Martian ejecta emplacement. *J. Geophys. Res.* **84**, 7669–7667 (1979).
20. Barlow, N. G. *et al.* Standardizing the nomenclature of Martian impact crater ejecta morphologies. *J. Geophys. Res.* **105**, 26733–26738 (2000).
21. Baker, V. R., Gulick, V. C. & Kargel, J. S. in *Resources of Near-Earth Space* (eds Lewis, J. S., Matthews, M. S. & Guerrieri, M. L.) 765–797 (Univ. Arizona Press, Tucson, 1993).
22. Chapman, M. G. *et al.* In *Environmental Effects on Volcanic Eruptions: From Deep Oceans to Deep Space* (eds Zimbelman, J. R. & Gregg, T. K. P.) 39–73 (Kluwer Academic, Plenum, New York, 2000).
23. Hartmann, W. K. & Berman, D. C. Elysium Planitia lava flows: crater count chronology and geological implications. *J. Geophys. Res.* **105**, 15011–15025 (2000).
24. Burr, D. & McEwen, A. S. in *Extremes of the Extremes* (eds Snorrason, A. & Finnssdottir, H. P.) Int. Assoc. Sci. Hydrol. Spec. Publ. (Int. Assoc. Sci. Hydrol., in the press).
25. Keszthelyi, L. & McEwen, A. S. Terrestrial analogs and thermal models for Martian flood lavas. *J. Geophys. Res.* **105**, 15027–15049 (2000).
26. Hodges, C. A. & Moore, H. J. The subglacial birth of Olympus Mons and its aureoles. *J. Geophys. Res.* **84**, 8061–8074 (1979).
27. Mouginis-Mark, P. J. in *Lunar and Planetary Science XXXIV* 1021–1022 (Lunar Planet. Inst., Houston, 1993).
28. Moore, J. G. *et al.* Prodigious submarine landslides on the Hawaiian Ridge. *J. Geophys. Res.* **94**, 17465–17484 (1989).
29. Pieri, D. C. *Geomorphology of Martian Valleys*. NASA Rep. No. TM-81979 (NASA, Washington DC, 1980).
30. Gulick, V. C. & Baker, V. R. Fluvial valleys and martian palaeoclimates. *Nature* **341**, 514–516 (1989).
31. Baker, V. R. *et al.* in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 493–522 (Univ. Arizona Press, Tucson, 1992).
32. Gulick, V. C. Origin of the valley networks on Mars: a hydrological perspective. *Geomorphology* **37**, 241–268 (2001).
33. Laity, J. E. & Malin, M. C. Sapping processes in the development of theater-headed valley networks in the Colorado Plateau. *Geol. Soc. Am. Bull.* **96**, 203–217 (1985).
34. Baker, V. R. *et al.* in *Ground-water Geomorphology: The Role of Subsurface Water in Earth-Surface Processes and Landforms* (eds Higgins, C. G. & Coates, D. R.) Geol. Soc. Am. Spec. Pap. 252, 235–265 (Geol. Soc. Am., Boulder, CO, 1990).
35. Carr, M. H. & Clow, G. D. Martian channels and valleys: their characteristics, distribution, and age. *Icarus* **48**, 91–117 (1981).
36. Scott, D. H., Dohm, J. M. & Rice, J. W. Jr Map showing channels and possible paleolake basins. US Geol. Surv. Misc. Invest. Ser. MAP I-2461 (1995).
37. Baker, V. R. & Partridge, J. Small Martian valleys: pristine and degraded morphology. *J. Geophys. Res.* **91**, 3561–3572 (1986).
38. Carr, M. H. & Chuang, F. C. Martian drainage densities. *J. Geophys. Res.* **102**, 9145–9152 (1997).
39. Carr, M. H. & Malin, M. C. Meter-scale characteristics of Martian channels and valleys. *Icarus* **146**, 366–386 (2000).
40. Gulick, V. C. Magmatic intrusions and a hydrothermal origin for fluvial valleys on Mars. *J. Geophys. Res.* **103**, 19365–19387 (1998).
41. Phillips, R. J. *et al.* Ancient geodynamics and global scale hydrology on Mars. *Science* **291**, 2587–2591 (2001).
42. Craddock, R. A. & Maxwell, T. A. Geomorphic evolution of the Martian highlands through ancient fluvial processes. *J. Geophys. Res.* **98**, 3453–3468 (1993).
43. Hynek, B. M. & Phillips, R. J. Evidence for extensive denudation of the martian highlands. *Geology* **29**, 407–410 (2001).
44. Golombek, M. P. & Bridges, N. T. Erosion rates on Mars and implications for climate change: constraints from the Pathfinder landing site. *J. Geophys. Res.* **105**, 1841–1853 (2000).
45. Carr, M. H. Retention of an atmosphere on early Mars. *J. Geophys. Res.* **104**, 21897–21909 (1999).
46. Baker, V. R. *et al.* Ancient oceans, ice sheets, and the hydrological cycle on Mars. *Nature* **352**, 589–594 (1991).
47. Baker, V. R. & Milton, D. J. Erosion by catastrophic floods on Mars and Earth. *Icarus* **23**, 27–41 (1974).
48. Baker, V. R. & Nummedal, D. *The Channeled Scabland* (NASA Planetary Geology Program, Washington DC, 1978).
49. Baker, V. R. *The Channels of Mars* (Univ. Texas Press, Austin, 1982).
50. Komar, P. D. Modes of sediment transport in channelized water flows with ramifications to the erosion of the Martian outflow channels. *Icarus* **42**, 317–329 (1980).
51. Baker, V. R. Erosional processes in channelized water flows on Mars. *J. Geophys. Res.* **84**, 7985–7993 (1979).
52. Nummedal, D. & Prior, D. B. Generation of Martian chaos and channels by debris flows. *Icarus* **45**, 77–86 (1981).
53. Lucchitta, B. K. Antarctic ice streams and outflow channels on Mars. *Geophys. Res. Lett.* **28**, 403–406 (2001).
54. Baker, V. R. in *Flood and Megaflood Deposits: Recent and Ancient Examples* (eds Martini, I. P., Baker, V. R. & Garzon, M.) Int. Assoc. Sedimentol. Spec. Publ. (Int. Assoc. Sedimentol., in the press).
55. Broecker, W. S. & Denton, G. H. The role of ocean-atmosphere reorganizations in glacial cycles. *Geochim. Cosmochim. Acta* **53**, 2465–2501 (1989).
56. Dohm, J. M. *et al.* System of gigantic valleys northwest of Tharsis, Mars: latent catastrophic flooding, northwest watershed, and implications for northern plains ocean. *Geophys. Res. Lett.* **27**, 3559–3562 (2000).
57. Nelson, D. M. & Greeley, R. Geology of Xanthe Terra outflow channels and the Mars Pathfinder landing site. *J. Geophys. Res.* **104**, 8653–8669 (1999).
58. Rotto, S. & Tanaka, K. L. Geologic/geomorphic map of the Chryse Planitia region of Mars. US Geol. Surv. Misc. Invest. Ser. MAP I-2441 (1995).
59. Ivanov, M. A. & Head, J. W. Chryse Planitia, Mars: topographic configuration, outflow channel continuity and sequence, and tests for hypothesized ancient bodies of water using Mars Orbiter Laser Altimeter (MOLA) data. *J. Geophys. Res.* **106**, 3275–3295 (2001).
60. Burr, D. M., McEwen, A. S. & Sakimoto, S. E. H. Recent aqueous floods from the Creberus Rupes, Mars. *Geophys. Res. Lett.* (submitted).
61. Mouginis-Mark, P. J. Recent water release in the Tharsis region of Mars. *Icarus* **84**, 362–373 (1990).
62. Carr, M. H. Channels and valleys on Mars: cold climate features formed as a result of a thickening cryosphere. *Planet. Space Sci.* **44**, 1411–1423 (1996).
63. Carr, M. H. Martian oceans, valleys and climate. *Astron. Geophys.* **41**, 3.20–3.26 (2000).
64. Head, J. W. III & Wilson, L. Mars: geologic setting of magma/H₂O interactions. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1215 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1215.pdf>> (2001).
65. Tanaka, K. L. & Chapman, M. G. The relation of catastrophic flooding of Mangala Valles, Mars, to faulting of Memnonia Fossae and Tharsis volcanism. *J. Geophys. Res.* **95**, 14315–14323 (1990).
66. Max, M. D. & Clifford, S. M. Initiation of Martian outflow channels: related to the dissociation of gas hydrate? *Geophys. Res. Lett.* **28**, 1787–1790 (2001).
67. Komatsu, G. *et al.* A chaotic terrain formation hypothesis: explosive outgas and outflow by dissociation of clathrate on Mars. *Lunar Planet. Sci. Conf. XXXI*, Abstr. 1434 <<http://www.lpi.usra.edu/meetings/lpsc2000/pdf/1434.pdf>> (2000).
68. Jöns, H. P. Late sedimentation and late sediments in the northern lowlands of Mars. *Lunar Planet. Sci.* **16**, 414–415 (1985).
69. Lucchitta, B. K., Ferguson, H. M. & Summers, C. Sedimentary deposits in the northern lowland plains, Mars. *J. Geophys. Res.* **91**, E116–E174 (1986).
70. Parker, T. J., Saunders, R. S. & Shcneeberger, D. M. Transitional morphology in the west Deuteronilus Mensae region of Mars: implications for modification of the lowland/upland boundary. *Icarus* **82**, 111–145 (1989).
71. Head, J. W. III *et al.* Possible ancient oceans on Mars: evidence from Mars Orbiter Laser Altimeter Data. *Science* **286**, 2134–2137.
72. Parker, T. J. *et al.* Coastal geomorphology of the Martian northern plains. *J. Geophys. Res.* **98**, 11061–11078 (1993).
73. Malin, M. C. & Edgett, K. S. Oceans or seas in the Martian northern lowlands: high-resolution imaging tests of proposed coastlines. *Geophys. Res. Lett.* **26**, 3049–3052 (1999).
74. Tanaka, K. L. Debris flow origin for the Simud/Tiu deposit on Mars. *J. Geophys. Res.* **104**, 8637–8652 (1999).

75. Kargel, J. S. *et al.* Evidence for continental glaciation in the Martian northern plains. *J. Geophys. Res.* **100**, 5351–5368 (1995).
76. Clifford, S. M. & Parker, T. J. The evolution of the Martian hydrosphere: implications for the fate of a primordial ocean and the current state of the northern plains. *Icarus* (in the press).
77. Cabrol, N. A. & Grin, E. A. Distribution, classification, and ages of Martian impact crater lakes. *Icarus* **142**, 160–172 (1999).
78. Cabrol, N. A. & Grin, E. A. The evolution of lacustrine environments on Mars: is Mars only hydrologically dormant? *Icarus* **149**, 291–328 (2001).
79. Ori, G. G., Marinangeli, L. & Baliva, A. Terraces and Gilbert-type deltas in crater lakes in Ismenius Lacus and Memnonia (Mars). *J. Geophys. Res.* **105**, 17629–17641 (2000).
80. Squyres, S. W. Urey Prize Lecture: water on Mars. *Icarus* **79**, 229–288 (1989).
81. Scott, D. H. & Chapman, M. G. Geologic and topographic maps of the Elysium paleolake basin, Mars. US Geol. Surv. Geol. Ser. MAP I-2397 (1995).
82. Cabrol, N. A. *et al.* Hydrogeologic evolution of Gale Crater and its relevance to the exobiological exploration of Mars. *Icarus* **139**, 235–245 (1999).
83. Malin, M. C. & Edgett, K. S. Sedimentary rocks of early Mars. *Science* **290**, 1927–1937 (2000).
84. Weitz, C. M. *et al.* The interior layered deposits of Valles Marineris: layering, erosional processes, and age relationships. *Lunar Planet. Sci.* **32**, Abstr. No. 1629 (2001).
85. Lucchitta, B. K. MOC images confirm layered deposits formed within Valles Marineris, Mars. *Lunar Planet. Sci.* **32**, Abstr. No. 1359 (2001).
86. Malin, M. C. & Edgett, K. S. Evidence for recent groundwater seepage and surface runoff on Mars. *Science* **288**, 2330–2335 (2000).
87. Costard, F. *et al.* Debris flows on Mars: Analogy with terrestrial periglacial environment and climatic implications. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1534 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1534.pdf>> (2001).
88. French, H. M. *The Periglacial Environment* (Longman, Harlow, 1996).
89. Musselwhite, D. S., Swindle, T. D. & Lunine, J. I. Liquid CO₂ breakout and the formation of recent small gullies on Mars. *Geophys. Res. Lett.* **28**, 1283–1286 (2001).
90. Hoffman, N. White Mars: a new model for Mars' surface and atmosphere based on CO₂. *Icarus* **146**, 326–342 (2000).
91. Wilson, L. Les relations entre les processus geomorphologique et le climat moderne comme méthode de paléoclimatologie. *Rev. Géogr. Physique Geol. Dynamique* **11**, 309–314 (1969).
92. Lucchitta, B. K. Ice sculpture in the Martian outflow channels. *J. Geophys. Res.* **87**, 9951–9973 (1982).
93. Costard, F. & Baker, V. R. Thermokarst landforms and processes in Ares Vallis, Mars. *Geomorphology* **37**, 287–301 (2001).
94. Kargel, J. S. & Strom, R. G. Ancient glaciation on Mars. *Geology* **20**, 3–7 (1992).
95. Head, J. W. III & Hallet, B. Origin of sinuous ridges in the Dorsa Argentea Formation: new observations and tests of the esker hypothesis. *Lunar Planet. Sci.* **32**, Abstr. No. 1373 (2001).
96. Head, J. W. III & Pratt, S. Extensive Hesperian-aged south polar ice sheet on Mars: evidence for massive melting and retreat, and lateral flow and ponding of meltwater. *J. Geophys. Res.* (in the press).
97. Mangold, N. & Allemand, P. Topographic analysis of features related to ice on Mars. *Geophys. Res. Lett.* **28**, 407–410 (2001).
98. Colaprete, A. & Jakosky, B. M. Ice flow and rock glaciers on Mars. *J. Geophys. Res.* **103**, 5897–5909 (1998).
99. Clark, D. H., Steig, E. J., Potter, N. Jr & Gillespie, A. R. Genetic variability of rock glaciers. *Geogr. Annls* **80A**, 175–182 (1998).
100. Barsch, D. *Rockglaciers* (Springer, Berlin, 1996).
101. Squyres, S. W., Clifford, S. M., Kuzmin, R. O., Zimbelman, J. R. & Costard, F. M. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 523–554 (Univ. Arizona Press, Tucson, 1992).
102. Ackert, R. P. Jr A rock glacier/debris-covered glacier at Galena Creek, Absaroka Mountains, Wyoming. *Geogr. Annls* **80A**, 267–276 (1998).
103. Carr, M. H. D/H on Mars: effects of floods, volcanism, impacts and polar processes. *Icarus* **87**, 210–227 (1990).
104. Mulvaney, R. R. *et al.* The transition from the last glacial period in inland and near-coastal Antarctica. *Geophys. Res. Lett.* **27**, 2673–2676 (2000).
105. Sugden, D. E. *et al.* Preservation of Miocene glacier ice in East Antarctica. *Nature* **376**, 412–414 (1995).
106. Crowell, J. C. *Pre-Mesozoic Ice Ages: Their Bearing on Understanding the Climate System* Geol. Soc. Am. Memoir 192 (Geol. Soc. Am., Boulder, CO, 1999).
107. Hoffman, P. F. *et al.* A Neoproterozoic snowball Earth. *Science* **281**, 1342–1346 (1998).
108. Williams, G. E. History of the Earth's obliquity. *Earth-Sci. Rev.* **34**, 1–45 (1993).
109. Hoffman, P. F. & Schrag, D. P. The Snowball Earth hypothesis: theory, observations, and tests. *Terra Nova* (in the press).
110. Hays, J. D., Imbrie, J. & Shackleton, N. J. Variations in the Earth's orbit: pacemaker of the ice ages. *Science* **194**, 1121–1132 (1976).
111. Raymo, M. E. & Ruddiman, W. F. Tectonic forcing of late Cenozoic climate. *Nature* **359**, 117–122 (1992).
112. Touma, J. & Wisdom, J. The chaotic obliquity of Mars. *Science* **259**, 1294–1297 (1993).
113. Laskar, J. & Robutel, P. The chaotic deliquity of the planets. *Nature* **361**, 608–612 (1993).
114. Jakosky, B. M., Henderson, B. G. & Mellon, M. T. Chaotic obliquity and the nature of Martian climate. *J. Geophys. Res.* **100**, 1579–1584 (1995).
115. Rothschild, L. J. & Mancinelli, R. L. Life in extreme environments. *Nature* **409**, 1092–1101 (2001).
116. Sleep, N. H., Zahnle, K. & Neuhoff, P. S. Initiation of clement surface conditions on the earliest Earth. *Proc. Natl Acad. Sci. USA* (in the press).
117. Max, M. D. & Clifford, S. M. The state, potential distribution, and biological implications of methane in the Martian crust. *J. Geophys. Res.* **105**, 4165–4171 (2000).
118. Pollack, J. B. *et al.* The case for a wet, warm climate on early Mars. *Icarus* **71**, 203–224 (1987).
119. Anderson, R. C. *et al.* Primary centers and secondary concentrations of tectonic activity through time in the western hemisphere of Mars. *J. Geophys. Res.* (in the press).
120. Gulick, V. C. *et al.* Effects and lifetimes of ocean-induced CO₂ pulses on Mars: implications for fluvial valley formation. *Icarus* **130**, 68–86 (1997).

Acknowledgements

I thank many colleagues for comments and discussion useful to this review, including R. C. Anderson, D. Burr, N. Cabrol, F. M. Costard, J. M. Dohm, J. C. Ferris, E. Grin, V. C. Gulick, T. M. Hare, W. K. Hartmann, J. S. Kargel, G. Komatsu, A. S. McEwen, G. G. Ori, J. W. Rice Jr, R. G. Strom, K. L. Tanaka and J. R. Zimbelman. The entire manuscript was reviewed by J. W. Head III and by D. E. Sugden. NASA provided partial support for the research.

Mars' volatile and climate history

Bruce M. Jakosky*† & Roger J. Phillips*‡

*Laboratory for Atmospheric and Space Physics, and †Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309-0392, USA (e-mail: bruce.jakosky@lasp.colorado.edu)

‡McDonnell Center for Space Science and Dept. of Earth and Planetary Sciences, Washington University, St. Louis, Missouri 63130, USA (email: phillips@wustl.wustl.edu)

There is substantial evidence that the martian volatile inventory and climate have changed markedly throughout the planet's history. Clues come from areas as disparate as the history and properties of the deep interior, the composition of the crust and regolith, the morphology of the surface, composition of the present-day atmosphere, and the nature of the interactions between the upper atmosphere and the solar wind. We piece together the relevant observations into a coherent view of the evolution of the martian climate, focusing in particular on the observations that provide the strongest constraints.

Unravelling its climate history is one of the main challenges in understanding the evolution of Mars. There is substantial evidence for change in the climate and the inventory of volatiles through time. This evidence points to an early environment in which water was either more stable or more abundant at the surface than it is today, and to processes of supply of volatiles to the atmosphere and their removal from it that are able to explain this shift in climate. Determining the forces behind the changing climate, particularly the relationship between the atmosphere, water at the surface, and water in the crust, is important for understanding the planet as a whole. Additionally, the potential for life to have existed in the past, or even today, is strongly governed by the occurrence of liquid water and its history over time.

The climate system of Mars is inherently complex, involving physical and chemical processes within the deep interior, the crust, the surface, the atmosphere, the upper atmosphere, and interactions with the solar wind (Fig. 1). Our goal is to integrate observations and measurements pertinent to each of these areas into a coherent view of martian climate and its evolution. We will focus on recent results from the Mars Global Surveyor mission and from ongoing analyses of the martian meteorites, and will emphasize observations that provide the strongest constraints rather than describing what is merely plausible.

The present-day atmosphere and climate

The martian atmosphere consists predominantly of CO₂, with a total atmospheric pressure averaging about 6 mbar (6 hPa), about 0.6% of the Earth's atmospheric pressure. The atmosphere also contains N₂ and Ar at a level of 1–3%, lesser amounts of H₂O, CO and O₂, and additional gases in trace amounts¹.

Water vapour is present in the atmosphere, with a typical partial pressure of 10⁻³ mbar. This corresponds to about 10⁻³ g H₂O residing above each square centimetre, equivalent to a condensed layer about 10 µm thick if it all precipitated onto the surface; the water content of the Earth's atmosphere, in contrast, is about 10⁴ times greater. The atmospheric water content varies in a manner that is consistent with seasonal supply and removal from the polar caps, exchange with water adsorbed in the near-surface regolith, and global transport by means of winds².

Temperatures average about 220 K globally, well below the melting temperature of ice and lower than the eutectic freezing temperature of most salt-rich brines³. Even though

surface temperatures rise above 273 K over large regions near the equator and near noon, liquid water still would not be stable. It could exist as a transient phase, but would quickly evaporate into the relatively dry atmosphere and eventually freeze out at the colder high latitudes^{2,3}.

Mars' north polar region is covered with CO₂ ice during winter, which sublimates away and leaves a residual summer-time deposit of water ice. The south cap is covered year-round by CO₂ frost, but almost certainly has water ice mixed in or beneath it². The axial obliquity, the tilt of the polar axis with respect to the normal to the orbital plane (currently 25.2°), varies on timescales of 10⁵–10⁶ years (refs 4, 5). On longer timescales, it is chaotic, and may have been as low as 0° or as high as 60° in the past few million years (ref. 6). At high obliquity, polar summertime temperatures may increase markedly, and substantial amounts of water ice may sublimate into the atmosphere. Exchange of this water between the north and south polar caps, modulated by the changing eccentricity and season of perihelion, is probably responsible for the formation and evolution of layered deposits in the polar regions^{2,4,7}.

Nature of the earliest atmosphere and climate

The observable surface record of Mars' geological history spans 4 billion years (Gyr). The oldest, most heavily cratered surfaces are thought to be about 4.0 Gyr old, and the youngest are possibly less than 100 million years (Myr) old^{8–11}. Evidence pertaining to the ancient climate is inferred from processes that shaped the surface during the Noachian epoch, which ended when the cratering rate declined dramatically between 3.8 and 3.5 Gyr ago^{9,11}. (Fig. 2 summarizes the history and timing of processes involved in martian climate evolution.) Although atmospheric gas undoubtedly was present before 4.0 Gyr, and was removed in part from the atmosphere by various processes¹², we focus here on processes that post-dated the onset of the visible geological record.

The ancient martian surfaces contain geological features that indicate that the early climate was different. Dendritic networks of valleys seem similar to those formed by runoff of surface water on the Earth, although the areal density of tributaries is typically lower¹³. There is debate about the relative roles of surface runoff, sapping by release of subsurface water, and discharge of water from hydrothermal systems in forming the valleys^{14–16}. However, there is general agreement that water must have flowed at the surface in order to form these features and that their dendritic character and

typically V-shaped cross-section requires a gradual rather than catastrophic formation process (ref. 17, and see review in this issue by Baker, pages 228–236).

Erosion rates in general were substantially greater during the Noachian. This can be seen easily on ancient impact craters. The largest craters and basins are severely degraded; ejecta deposits, crater rims and central peaks have all been removed, and a paucity of craters smaller than about 15-km diameter suggests that they have been removed in their entirety^{18,19}. Some partially degraded craters have fluting and scalloping along the interior rim that suggests erosion by flowing surface water^{19,20}. Estimated erosion rates were more than 1,000 times larger than in subsequent epochs, and approach values appropriate for drier regions on Earth^{17,21}.

Additional evidence for a shift to a colder, drier climate at the end of the Noachian is provided by the initiation of U-shaped valley forms at the downstream ends of V-shaped valleys, suggesting an evolution of valley-forming erosional mechanisms from water-related to ice-related²². This period of U-shaped valley formation, and valley network evolution in general, ended rapidly at the end of the Noachian or early in the Hesperian^{16,22}.

Together, these features strongly suggest that liquid water was present during the Noachian and flowed over the surface, and that the climate must have been such that water was either more stable or more abundant at the surface than it is at present. Such a climate is thought to have required warmer surface temperatures produced by substantial greenhouse warming²³. Although the climate must have been 'warmer' and 'wetter', there is no consensus as to what temperatures would have been required, what the greenhouse gas would have been, or how much water would have to have been at the surface or in the atmosphere²⁴. If CO₂ were the greenhouse gas, up to several bars pressure would have been required to produce the necessary warming^{25–27}. A warmer climate is indicated even if valley network formation involved subsurface hydrothermal processes, as flow at the surface and gradual erosion requires a warmer climate independent of the original source of the liquid water.

Connections between ancient climate and early geology

Other events also were taking place during the Noachian period that would have affected climate. The planet was being bombarded by impacting asteroids or planetesimals, it was creating its main division into a southern highlands and a northern lowlands (which is reflected today as the ancient and younger terrains, respectively), and volcanism was forming the major Tharsis province on which sits many of the large volcanoes.

The formation of Tharsis was recognized only recently as occurring largely in the Noachian^{28,29}. The heavy loading of the lithosphere, due to the dominantly magmatic formation of the Tharsis rise, resulted in a global warping of the surface²⁸. Many of the ancient valley networks are seen to preferentially follow the slopes that resulted from the formation of Tharsis or, in at least one location (Margaritifer Sinus), are of the same age as those that do. This indicates that emplacement of Tharsis must have been nearly complete while the largely Noachian valley networks were still forming. Although much of the surface in Tharsis is sparsely cratered and therefore relatively young, indicating that resurfacing of Tharsis continued throughout martian history, the bulk of the volume of Tharsis must be old²⁸.

The volcanic magma from which Tharsis formed probably contained substantial quantities of both water and CO₂ that would have been released, providing input of gases to the atmosphere and possibly contributing to an early, thicker atmosphere. Geochemical analysis of the martian meteorites, for example, suggests a water content of as much as 1.8% by weight³⁰. Thus, the timing of valley network formation may be more than coincidental. Tharsis volcanism may have supplied gases that helped maintain a climate conducive to weathering and erosion, and the cessation of volcanism may have allowed other processes to begin removal of much of the atmosphere (see below)²⁸.

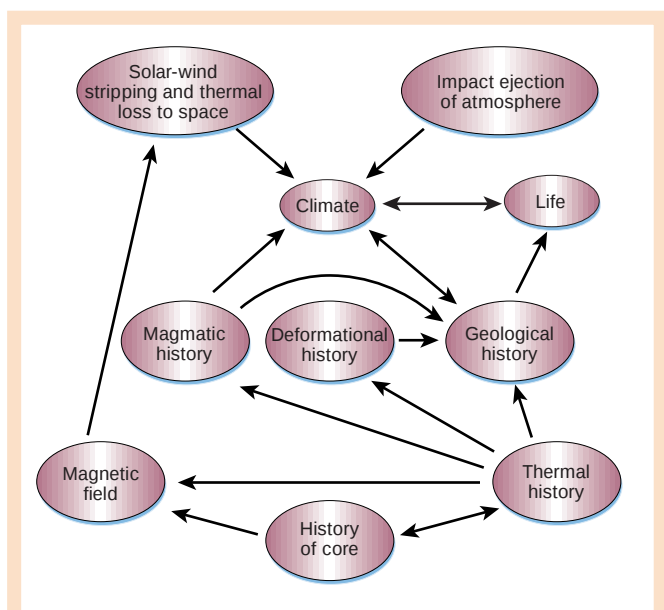


Figure 1 Schematic diagram showing the interconnected nature of the martian climate system and the relationship between climate and various processes from the deep interior to the upper atmosphere. Arrows indicate how one region affects another, with the arrow pointing towards the processes or area being affected; two-headed arrows indicate mutual interactions.

There is another compelling connection between Tharsis and the history of water. The lithospheric deformation due to the weight of Tharsis created a depression or trough encircling it at a radius of about 5,000 km (Fig. 3a). Most of the large-scale geological features related to surface water are concentrated in this trough²⁸. This includes the large-scale drainages that flow northward from the Argyre basin (at 50° S latitude) and the catastrophic flood channels that emanate from the eastern end of Valles Marineris and flow north into the Chryse basin and into the northern lowlands. Argyre, for example, seems to have been filled with water, which subsequently overflowed the rim and flowed northward³¹. Although the catastrophic flooding occurred episodically throughout martian history³², the source and flow regions concentrate in the Tharsis trough. This relationship suggests a long-lasting connection between the geological and geophysical history of Mars and the release, availability and geological effects of liquid water. But some anomalies exist, such as the paucity of similar features in the western branch of the circum-Tharsis trough and the ill-defined nature of processes, other than simple structural control of water pathways, that would provide a causal connection between the trough and the release of crustal water.

Finally, we note an apparent correlation between the regions on which crustal remanent magnetic fields are imprinted and areas on which valley networks are identified (Fig. 4)³³. The magnetic anomalies presumably retain an imprint of a global-scale intrinsic magnetic field from an early time in martian history (see below, ref. 34, and the review in this issue by Stevenson, pages 214–219). Exceptions to the correlation include the large-scale magnetic anomalies occurring in discrete bands in the high southern latitudes³⁵, where valley networks may have never formed because of climate constraints, and Margaritifer Sinus, where stripping of the surface may have removed the valleys without removing enough material to eliminate the magnetic anomalies³⁶. If genuine, this correlation may reflect a coincidence in timing of the formation of both the valleys and the magnetic anomalies or in the geographic location of subsequent alteration processes. Alternatively, it may involve a cause-and-effect relationship, in which valleys could represent the surface manifestation of hydrothermal systems driven by subsurface volcanism that thermally induced the creation of the localized remanent fields³⁷.

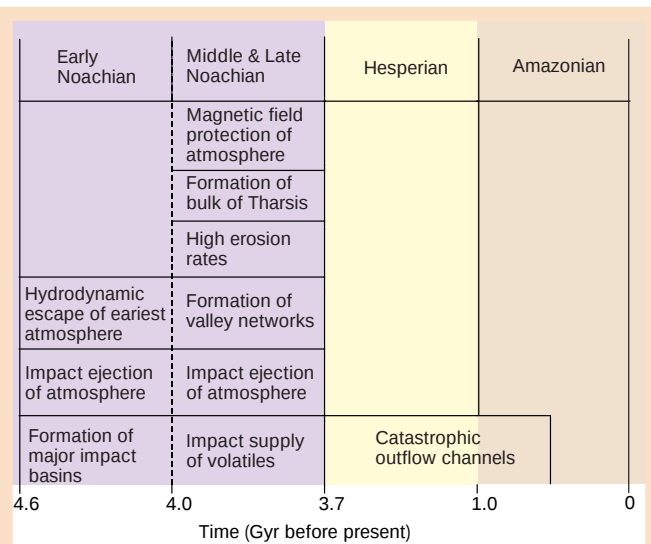


Figure 2 Schematic diagram showing the time history of the martian volatile system. The 4.6-Gyr history of Mars is divided up into the major geological epochs that have been identified, and processes that occur within each epoch are indicated. Note that the timeline is not a linear scale. Some climate-related processes are not shown (including recent polar-cap behaviour and the occurrence of possible crater lakes). Absolute ages marking the boundary between epochs are estimated based on the recent revision of cratering chronology¹¹.

However, as neither the mechanism of formation of the magnetic anomalies nor of their removal (if they ever were distributed globally) is understood, such connections are speculative at best.

Evolution of the Noachian atmosphere

What processes might have caused the inferred changes in Noachian climate, what sinks were available for atmospheric gases during this epoch, and can we infer the degree to which each process acted? These are especially important issues, as the output of the Sun was some 30% less 4 Gyr ago than it is today³⁸, making the maintenance of an early warmer climate more difficult^{39,40}.

Impact of large asteroids through the Noachian epoch would have ejected gas in the atmosphere to space⁴¹. Although inherently capable of removing large amounts of atmosphere, only those impacts occurring since the time of the onset of the geological record would have resulted in climate change that can be inferred from that record⁴². We know how many impacts occurred, based on the number of large craters and impact basins seen on the ancient terrain, and can readily extrapolate to impacts on areas subsequently buried⁴³. Combined, impacts are likely to have ejected about 50–90% of the atmosphere present in the early epoch⁴². This fraction probably is an upper limit, as some volatiles (such as water) might not have resided exclusively in the atmosphere at this time or might have been outgassed late in the epoch.

Recent measurements of surface elevation have allowed identification of previously unknown impact basins buried below a relatively thin veneer in both the southern highlands and the northern plains⁴³. The buried craters in the northern plains have similar abundance to the craters in the southern highlands, and justify the extrapolation of southern-hemisphere craters to a global inventory. The buried southern-highlands craters predate the geological evidence pertaining to the early climate, so those impacts would not have contributed to the changes in climate and should not be counted in summing up atmospheric loss.

Impacts of volatile-rich objects can supply new volatiles to Mars as well as remove them from the atmosphere⁴⁴. In particular, the analysis of isotope ratios of heavy noble gases in the atmospheres of Earth and Mars suggests that much of the planets' volatiles could have been supplied by comets⁴⁵. But the relative roles of supply and removal of volatiles by impacts is unknown.

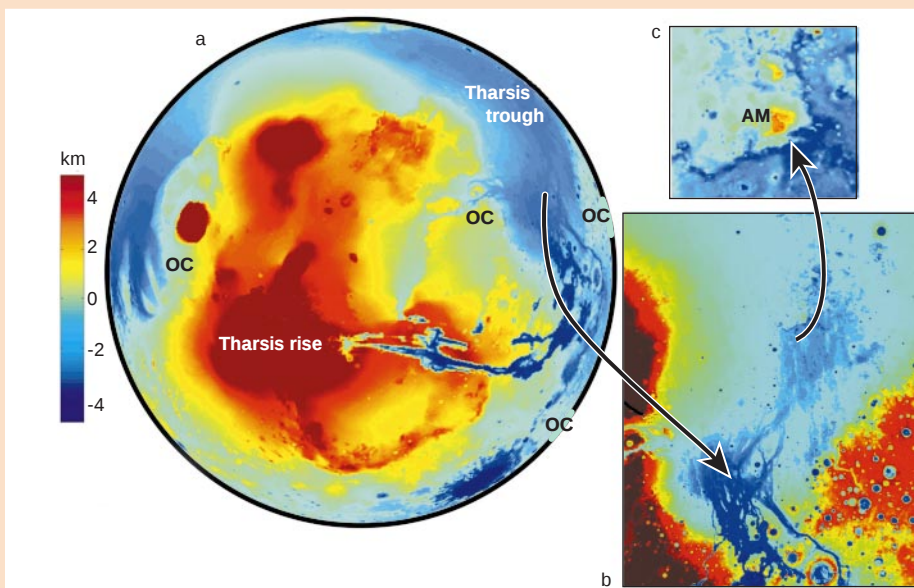
The impingement of the solar wind can strip gases directly from the upper atmosphere. Both atom-on-atom collisions that eject individual molecules (pick-up-ion sputtering) and hydrodynamic collisions that could strip off large volumes of gas *en masse* would have been important^{46,47}. But theoretical models of these processes are exceedingly uncertain even at the present epoch, and extrapolating to past epochs when the solar wind and solar ultraviolet radiation were more intense⁴⁸ compounds the uncertainties. However, two observations of the present-day atmosphere indicate that these processes were significant. First, *in situ* measurements of the energy spectrum of ionospheric electrons distinguish between ionization of gases in the upper atmosphere and those from the incoming solar wind. The morphology of discrete regions of solar-wind versus planetary electrons requires that large masses of upper atmosphere are being stripped off *en masse*⁴⁷. Second, isotopic measurements indicate that such stripping has occurred and has been significant. The lighter isotopes of each atom are enriched relative to the heavier ones at the top of the atmosphere by diffusive separation (for example, ¹²C relative to ¹³C, ¹⁴N to ¹⁵N, or ³⁶Ar to ³⁸Ar), and will be preferentially removed to space. Consequently, the gas remaining in the atmosphere becomes enriched in the heavier isotope⁴⁹. This enrichment pattern is observed consistently in the modern atmosphere. For example, the ratio ³⁸Ar/³⁶Ar is 30% greater on Mars than elsewhere in the Solar System, and nitrogen and carbon also show enrichments (Table 1). This enrichment indicates loss of between about 50 and 90% of the atmospheric species to space, with the actual loss probably tending towards the higher values^{50,51}.

In addition, compared to terrestrial values, martian water shows a fivefold enrichment of D relative to H (refs 52–54, and see Table 1). The lighter H atom can escape thermally to space more readily than D, so the considerable enrichment of observed D requires the loss of substantial quantities of hydrogen to space. The hydrogen comes from water, with the oxygen from water probably also being lost to space⁵⁵. The inferred water loss depends on the relative supply rates of H and D to the upper atmosphere, their relative escape rates, and the initial D/H ratio on Mars⁵⁶. Recent analysis of martian meteorites suggests that the initial D/H value on Mars was twice the terrestrial value⁵⁷; in addition, there has been a spectroscopic detection of upper-atmospheric D and a revision of atmospheric reaction rates^{54,58}. Although the initial interpretation was that about 90% of martian water had been lost to space⁵⁶, the current understanding suggests loss of about two-thirds of the water⁵⁹.

The observed isotopic enrichments indicate the fraction of each species that has been lost, but not the total amount lost. Atmospheric gases exchange with the regolith and polar caps, so it actually represents the fraction lost from the combined atmospheric plus non-atmospheric reservoir. Additionally, different portions of the polar ice, for example, may exchange with the atmosphere on different timescales⁷, and outgassing of juvenile gas at later times will affect the isotope ratios⁵¹. The absolute loss rates to space are uncertain today and may have varied through time⁶⁰, so although we can be confident that loss of substantial volatiles to space has occurred, we are uncertain as to how much has been lost, where it came from, or exactly how much has been left behind. (Although the losses of H and O are related, and one would expect the isotopic fractionation of D/H to be related to that of ¹⁸O/¹⁶O, oxygen has an additional reservoir with which it can exchange — minerals in the crust — so that the connection is likely to be complex.)

Whereas stripping by the solar wind of upper-atmospheric gases will change the isotope ratios, removal by impact will not. Impacts remove the gas from the entire atmospheric column, and thereby remove all isotopes with equal efficiency. Thus, the fractions of gas lost by these two processes add together. For example, if each indicates loss of 90%, these represent a different 90% and they sum up to a net loss of 99% of the atmospheric gas. Additionally, the two loss processes probably operated at different times, with impact erosion probably being most important early in the Noachian when impact rates were highest, and solar-wind stripping occurring late in the Noachian or into the Hesperian.

Figure 3 Topography⁹⁸ of Tharsis rise and Tharsis trough²⁸. **a**, Topography, which has been saturated at ± 5 km, with pole-to-pole slope⁹⁸ removed. Image is centred on 260° E longitude and view is from 10° north of equator. Outflow channel locations are marked by 'OC'. **b**, Channel detail in Chryse and Acidalia Planitiae (0°–60° N, 300°–0° E, Mercator projection; elevation range is –3.9 km to 0 km). **c**, The inlier Acidalia Mensa ('AM'), whose southern and eastern boundaries are channels (Mercator projection; elevation range is –3.0 km to –1.5 km). Eastern channel extends northwards of 50° N.



CO₂ can be removed from the atmosphere in the presence of liquid water to form carbonate minerals on the surface or in the subsurface²³. This process occurs on Earth, where gaseous CO₂ dissolves in the oceans, combines with calcium ions weathered from the continents, and forms deposits of calcite in limestone. If SO₂ was an early greenhouse gas, it might have formed crustal sulphate minerals. Carbonates have been found in the martian meteorites and have been shown to be indigenous to Mars^{61–63}. However, carbonate and sulphate minerals have not been detected in sufficient quantities that, if released in gaseous form, they could provide enough greenhouse warming to explain the early environment^{64,65}. The CO₂ from a several-bar atmosphere would form a global equivalent layer of carbonates perhaps a hundred metres thick; were this distributed throughout the entire volume of the crust, perhaps carried there by circulating groundwater, it might not be detectable at the surface. Thus, the extent to which carbonate minerals could be a sink for gases from an early atmosphere remains uncertain.

But processes do exist that together can account for loss of an early thicker atmosphere. Theoretical models of the timing of volatile loss suggest that these processes can account quantitatively for the inferred change in climate, but they are not unique⁴². Two measurements, however, provide supporting evidence.

First, gas trapped in the martian meteorite ALH84001 may represent a direct sample of the ancient atmosphere⁶⁶. Measurements of nitrogen and argon isotopes indicate that this gas is essentially unfractionated, meaning that most of the isotope-fractionating loss to space through solar-wind stripping had not yet occurred^{66,67}. Isotopic measurements of the gas Xe indicate that it was trapped into the rock 3.9 Gyr ago⁶⁸. If the Xe and lighter gases were incorporated into the rock at the same time and represent a sample of the martian atmosphere from that epoch, they provide a key time constraint on atmospheric loss (the bulk of the loss to space must have post-dated 3.9 Gyr). However, although the gases most plausibly came from the atmosphere and were incorporated into the rock at that time, there are concerns related to their uncertain carrier within the meteorite and to the presumed non-juvenile nature of the gases, such that the interpretation is not unique.

Second, measurements of localized remanent magnetic anomalies detected on Mars today³⁴ provide information on the timing of the shut-off of an intrinsic magnetic field. A substantial global magnetic field would cause the solar wind to stand off from the planet, limiting its ability to strip off the atmosphere⁶⁹. In addition, local magnetic anomalies each have the ability to protect the local

atmosphere, such that the atmosphere of a planet covered entirely by such anomalies also would be relatively well protected⁴⁷. Thus, the shut-off of an intrinsic magnetic field and the 'erasure' of local remanent magnetic anomalies in the crust, if they ever were distributed globally, would have allowed the turn-on of the stripping of the atmosphere by the solar wind.

The history of the magnetic field and the timing of its shut-off is extremely uncertain (see the contrasting views of refs 34, 70, and the review in this issue by Stevenson, pages 214–219). Most of the remanent magnetic anomalies occur in the ancient terrain, indicating that Mars had a substantial intrinsic magnetic field and that it turned off relatively early in martian history³⁴. Suggestion of a late turn-on of the global magnetic field⁷⁰ is at odds with magnetization found in carbonate globules in ALH84001, which indicates that a martian geodynamo was active 4 Gyr ago or earlier⁷¹. A few younger areas also have remanent fields, although there is a general absence of magnetic anomalies associated with Hesperian and Amazonian volcanics, adding to the confusion. It is clear, however, that an intrinsic magnetic field was most strongly connected to the Noachian epoch, and very much less so to later epochs. It may not be possible to resolve the timing issues more clearly, as it is uncertain what the carrier of the remanent magnetic fields is, whether remanent anomalies were originally distributed globally, and, if they were, what process would have erased them.

If the turn-off of the magnetic field and the erasure of localized magnetic anomalies allowed the solar wind to begin to strip species out of the atmosphere, then this atmospheric loss would have begun during the Noachian or early Hesperian. The turn-on of atmospheric loss would have been complete only when a substantial fraction of local anomalies had been destroyed. This may not have happened prior to the substantial volcanism associated with Tharsis and the subsequent Hesperian era.

Together, the isotopic and magnetic-field evidence indicate, but do not prove, that the bulk of solar-wind stripping to space occurred subsequent to 3.9 Gyr, and that the turn-on of loss occurred sometime between the Middle Noachian and the Early Hesperian. Loss by solar-wind stripping seems to have been roughly contemporaneous with the change in climate inferred from geology, with the main changes in climate coinciding with the end of the Noachian.

Evidence for liquid water in later epochs

There is substantial evidence that liquid water has been present within the martian crust up to the present. Evidence comes from geological features seen on the surface as well as from geochemical analyses.

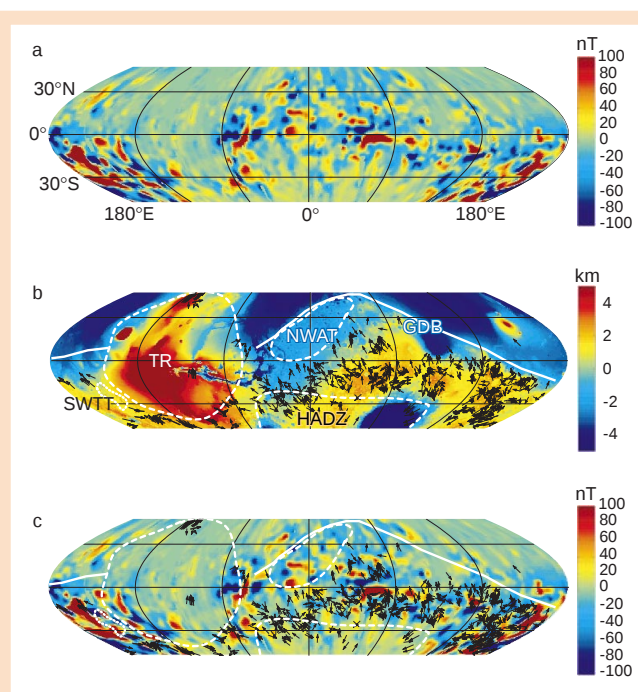


Figure 4 Comparison of martian magnetic field, topography and valley network locations. **a**, Radial component of the martian magnetic field at an altitude of 200 km (after ref. 106). Values saturated at ± 100 nanotesla (nT). Sinusoidal projection to $\pm 47.5^\circ$ latitude. **b**, Topographic map of Mars (after ref. 98), saturated at ± 5 km elevation, and locations of valley network systems¹³. Database is limited to latitudes less than $\pm 47.5^\circ$, which is the approximate limit of valley network distribution. Features indicated are the Tharsis rise (TR), global dichotomy boundary (GDB), northwestern Arabia Terra (NWAT), southwestern Tharsis trough (SWTT) and Hellas-Argyre disruption zone (HADZ). TR, NWAT, SWTT and the area north of GDB are regions of valley network exclusion by erosion and/or deposition and cannot be judged for the level of correlation of valley networks with anomalies in the remanent magnetic field. Likewise, TR, HADZ and the area north of GDB are regions of magnetic field exclusion by the presence of post-main-field crustal ages or crustal disruption³⁴. **c**, Valley networks and features in **b** superposed on magnetic anomaly map.

Catastrophic outflow channels provide compelling evidence that liquid water was released from within the crust, flowed over the surface, and drained into the northern lowlands (see refs 17, 72, and the review in this issue by Baker, pages 228–236). As these floods involved large quantities of water released catastrophically, with water able to flow substantial distances before freezing, they could occur even in the present cold climate¹⁷. They do indicate, though, that liquid water must have been present within the crust. Other eroding agents have been suggested, including liquid CO_2 , SO_2 , volcanic lava, ice, debris flows and the wind^{73–75}. In particular, the similarity of some of the martian channels to volcanic channels, and the ability of the wind to weather and erode substantial amounts of material, have recently been highlighted (C. B. Leovy and J. C. Armstrong, manuscript in preparation), along with the difficulty of reaching unique conclusions about the eroding fluid from morphology alone (see review by Baker). However, water or water-filled debris flows⁷⁵ remain the most plausible candidates and require the fewest extrapolations (see review by Baker). Although the observed floods seem to have drained only about 10% of the surface, connected largely to the Tharsis trough, there is no reason to think that crustal water was not distributed globally⁷⁶.

The martian meteorites contain weathering products, produced when liquid water was present, filling cracks and voids in the rock. These include carbonate deposits at levels of several per cent in ALH84001, as well as trace amounts of carbonates and the mineral iddingsite (which forms from weathering of basalt in the presence of liquid water) in several of the other meteorites^{61–63}. Detailed micro-

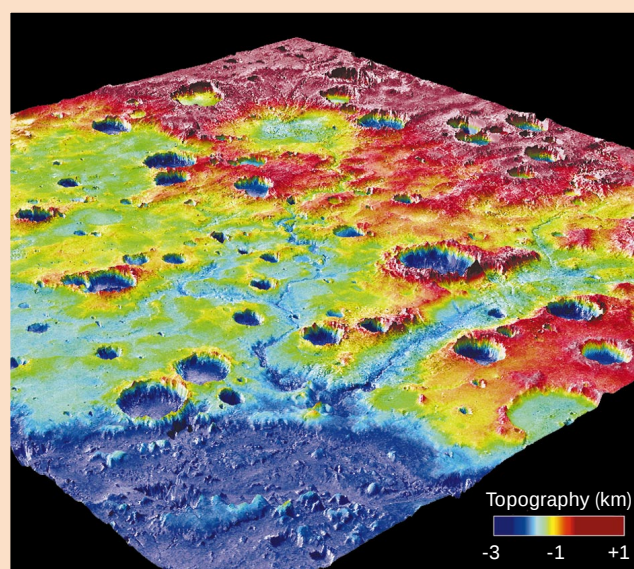


Figure 5 View of topography (from ref. 98) looking towards the southeast in Margaritifer Sinus region. Area is $30^\circ \times 30^\circ$ in size and there is $\times 12$ vertical exaggeration in topography. Pole-to-pole slope has been removed in topography to emphasize regional variations. The prominent depression in the centre of the image is Parana basin, a possible site of lake sediments¹⁰⁷; it is drained by a prominent channel (Loire Vallis) that descends northwesterly into the Tharsis trough (foreground). Higher elevations (red shades) are Early- and Middle-Noachian plateau materials. Yellow, green and blue/grey shades of lower elevations represent widespread erosion events in the Late Noachian that removed $\sim 1.5 \times 10^6 \text{ km}^3$ of Early/Middle-Noachian material from this region³⁶. Erosional remnants of older material occur in isolated mesas. (Image courtesy of B. Hynek.)

stratigraphy shows that the deposits were present before the rocks were ejected from the martian subsurface, providing direct evidence that liquid water circulated through the martian crust^{61,62}.

Indirect geochemical evidence for liquid water in the crust comes from measurements of various isotope ratios within these weathering products. They contain enhanced D/H, $^{13}\text{C}/^{12}\text{C}$, $^{15}\text{N}/^{14}\text{N}$ and $^{38}\text{Ar}/^{36}\text{Ar}$ ratios⁶². Enhancement of each of these ratios is best explained as having resulted from atmospheric processes involving preferential escape of the lighter isotopes to space (as discussed earlier). This requires that these gases once resided in the atmosphere, were left behind as a residuum of loss to space, and were subsequently incorporated into the crust. Circulation of groundwater between the surface and the crust provides the best means for exchange of these gases, again suggesting the presence of liquid water^{62,77,78}.

While the carbonates in ALH84001 were probably precipitated from liquid water about 3.9 Gyr ago, water also must have been present much more recently. The other martian meteorites are all much younger rocks, having crystallized between about 170 Myr and 1.3 Gyr, and also contain weathering deposits^{79,80}. Thus, crustal water must have been present within the past billion years. Radiogenic dates for the weathering products in the Lafayette meteorite indicate deposition around 650 Myr ago or even more recently⁸¹.

Several locations on the surface show spectroscopic evidence for the presence of coarse-grained haematite, probably formed in association with liquid water⁸². The haematite may have been deposited from water released from volcanic intrusions or driven by subsurface hydrothermal or aqueous systems⁸², although some could have formed by volcanic processes not requiring water⁸³. One of the occurrences is in Aram Chaos, a feature attributed to the release of subsurface water and not connected to volcanism. A second, larger location is in Western Terra Meridiani, located at about 0° latitude and 0° longitude, and is thought to have formed at depth and been exposed subsequently by erosion⁸².

Liquid water also may have been present within the near-surface crust very recently, based on the presence of pristine gullies on the exposed walls of impact craters and valleys⁸⁴. These are identical in size, shape and appearance to gullies on Earth carved by liquid water seeping from aquifers on exposed scarps. Liquid CO₂ has been suggested as a possible eroding agent, based on the similarity of the depth of seeping to that at which the overburden pressure equals the CO₂ liquefaction pressure⁸⁵. However, the absence of a viable charge or recharge mechanism for liquid CO₂ and the inability of CO₂ to discharge as a liquid under martian conditions preclude its role as a significant erosive agent⁸⁶. Rather, liquid water is much more plausible geologically (ref. 84, and see review by Baker, pages 228–236). Recent calculations of the stability of liquid water in the crust, and of the ability of cyclical oscillations in temperature to freeze and release water, suggest that this is a viable mechanism⁸⁷. There is no means to uniquely determine the age of the features, but they are unweathered and gully debris overlies features such as sand dunes that are themselves thought to be extremely young. If these interpretations are correct, liquid water has been present within a few hundred meters of the surface within the past few million years, and may reside there today⁸⁴.

There also is geological evidence for liquid water at the surface, possibly during the later epochs of martian history. Most intriguing is the evidence for standing lakes within impact craters⁸⁸ and the Valles Marineris canyon system¹⁷. Potential crater lakes are identified by the presence of channels flowing into and/or out of the craters, providing a source or sink for water. Some deposits within craters have an appearance similar to deltas, sedimentary terraces and shorelines, as might be formed by flowing water. Although most of the craters themselves are relatively old, the age of the lakebed deposits is uncertain. There are few craters on the lakebed deposits, which indicates either that they are very young⁸⁸ or that they were buried for long periods and were exhumed very recently⁸⁹; stratigraphic relationships of layered sedimentary deposits within Valles Marineris suggest an old age despite the absence of craters⁸⁹. Well-defined layering within these deposits supports the idea of standing water, although layered lakebed deposits cannot be distinguished uniquely from windblown sediments⁸⁹. But even if these sediments were windblown, they would have required either liquid water at least in trace amounts or water ice that might be stable at other epochs in order to cement the grains together to form coherent layers.

The catastrophic outflow channels all drain into the northern lowlands¹⁷. Possible shoreline features have been identified, suggesting that the water may have accumulated in the lowlands to form a large, long-lived body of water (that is, an ocean)^{90,91}. The crudely constant elevation of the innermost of the two main proposed shorelines⁹² and the extreme smoothness of the lowlands are held to be consistent with the presence of an ocean⁹³. However, the proposed shoreline features are not visible in recent high-resolution images⁹⁴ and thus cannot be ascribed with certainty to wave action at an ocean boundary. Features described originally as 'high stands' or shorelines created by a retreating ocean⁹² are now recognized as being wrinkle ridges of tectonic derivation⁹⁵. These ridges underlie the Vastitas Borealis formation, a deposit that may in part be sedimentary in origin⁹⁶.

The smoothness of the lowlands^{97,98} may have resulted from fluvially transported sediments associated with channel emplacement during the Hesperian⁹⁹; such a process allows but does not require the presence of an ocean. Erosion of material in the Margaritifer Sinus region and an adjacent area in northwestern Arabia Terra may have provided a significant source of sediments³⁶. About a kilometre of material seems to have been stripped away during the Late Noachian, and geological evidence points to erosion by liquid water (Fig. 5). The resulting debris could have filled the northern plains to a depth of about 100 m north of 30° N (ref. 36). The corresponding equivalent depth of water necessary to transport this sediment would have been several times the sediment volume¹⁰⁰. Whether or not this resulted in a Late Noachian ocean would have depended on the relative rates of

Table 1 Martian isotope ratios and atmospheric loss*

Isotope ratio	Measured value†	Amount lost to space (%)‡
D/H	5	~60–74
³⁶ Ar/ ³⁹ Ar	1.3	~50–90
¹³ C/ ¹² C	1.05–1.07	~50–90
¹⁵ N/ ¹⁴ N	1.7	~90
¹⁸ O/ ¹⁶ O	1.025	~25–50

*Values taken from refs 57–59, 62, 77 and 78, and references therein.

†Value estimated, observed or derived for martian atmosphere relative to terrestrial.

‡Calculated assuming Rayleigh fractionation. D/H range includes uncertainty in escape processes. Other ranges are based on uncertain timing of outgassing relative to escape.

erosion and of water removal by ground infiltration, evaporation and so on. Linear gravity anomalies trending northward from Chryse Planitia may be indicative of early channels buried in the northern plains, analogous to the main catastrophic flood channels seen to the east of Valles Marineris¹⁰¹. These could be Noachian-era pathways that moved water and sediment from the southern highlands to the northern lowlands.

In addition, the Hesperian outflow flood channels can be traced northward into Acidalia Planitia. There, they converge into a basin (not presently closed) with an area $\sim 6 \times 10^5$ km², south of the high-standing Noachian inlier Acidalia Mensa at 45° N latitude (Fig. 3b). Acidalia Mensa itself has erosional boundaries on the south and east that are channels (Fig. 3c), and may at one time have formed a natural barrier to water flow. This region of channel convergence coincides closely with one of two main occurrences of kilometre-scale polygonally fractured terrain¹⁰². The origin of polygonally fractured terrain is controversial^{103,104}, but many of the processes proposed involve water. Thus, it is likely that water (and thus sediments) were carried far into the northern lowlands during the Hesperian, perhaps forming lakes. The region of channel convergence in Acidalia Planitia may mark the largest possible extent of any standing bodies of surface water. If there ever was a Hesperian northern 'ocean', it may have been more of a large lake about the size of the Caspian Sea ($\sim 4 \times 10^5$ km²).

Synthesis of the observations and interpretations

It is clear that there are a number of strong constraints on the history of water on Mars. They tell us that water was present at the surface early in the planet's history, and within the crust throughout time. They also allow us to construct a self-consistent scenario of the history of martian volatiles and climate, although such a scenario is not unique.

The history of martian volatiles involved the following (see Fig. 2).

1. Most of the earliest atmosphere of Mars was lost during the Early Noachian by impact erosion and hydrodynamic escape.
2. A secondary atmosphere was created by water and CO₂ released to the atmosphere as a direct result of Tharsis volcanism, and this may have had a strong influence on climate. It is likely that volatiles were also released by non-Tharsis Noachian volcanism presumed to have been responsible for forming the ancient highland crust.
3. Water and CO₂ were lost from the surface and atmosphere system to space, to the polar caps and to carbonate deposits within the crust. There is compelling evidence for the existence of each of these sinks, as described above, although it is not possible at this time to determine uniquely the relative or absolute importance of each.
4. There is a coincidence in the timing of major events in martian history. The decrease in the impact rate at the end of heavy bombardment, the formation of the bulk of the Tharsis construct by magmatism, the decline in the intensity or existence of an intrinsic magnetic field, the change in climate inferred from the morphological characteristics of the surface, and the loss of substantial volatiles to space all occurred at nearly the same time and marked the end of the Noachian epoch about 3.7 Gyr ago. Many of these events are likely to have been causally connected to each other, although some of the similarity in timing may be coincidental. However, the loss of atmospheric protection by the shutdown of the global magnetic field

and the decline in the rate of Tharsis volcanism at the end of the Noachian probably both were instrumental in the major shift in climate as inferred from the geological record.

5. There has been a reservoir of crustal liquid water throughout martian history, continuing possibly to the present. This reservoir has exchanged, at various times and to various degrees, with water at the surface and in the atmosphere. We have little information as to how much water is contained in this reservoir today, but it is likely to be both substantial, globally distributed, and readily accessible to the surface in some places.

New measurements from both planned and potential missions will allow us both to test the above hypotheses and to further constrain the processes that have taken place (see the article in this issue by Carr and Garvin, pages 250–253). Although the list of relevant observations is long¹⁰⁵, some measurements will be key to improving our understanding. These include determining the present-day isotope ratios of the climate-related species (H, C, O and N) and the noble gases; sampling rocks over a broad age spectrum and determining the abundance of weathering products and the isotope ratios of gases contained within them; continuing the search for evidence for trapped carbonate or sulphate minerals within the crust, especially in rocks previously underground and now exposed at the surface; and continuing to explore the morphology of the surface for evidence of climate-related processes.

Discoveries made in the past decade have had a tremendous impact on our understanding of the history of martian water, volatiles and climate. Analysis of the martian meteorites and the measurements made from the Mars Global Surveyor spacecraft in particular have been key to progress in this area. Unravelling the history of martian volatiles and climate will be central to addressing in a meaningful way the potential for martian life and to interpreting results obtained specifically to look for evidence of present or past life. □

1. Owen, T. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder C. W. & Matthews, M. S.) 818–834 (Univ. Arizona Press, Tucson, 1992).
2. Jakosky, B. M. & Haberle, R. M. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder C. W. & Matthews, M. S.) 969–1016 (Univ. Arizona Press, Tucson, 1992).
3. Brass, G. W. Stability of brines on Mars. *Icarus* **42**, 20–28 (1980).
4. Kieffer, H. H. & Zent, A. P. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder C. W. & Matthews, M. S.) 1180–1218 (Univ. Arizona Press, Tucson, 1992).
5. Ward, W. R. Climatic variations on Mars. I. Astronomical theory of insolation. *J. Geophys. Res.* **79**, 3375–3386 (1974).
6. Touma, J. & Wisdom, J. The chaotic obliquity of Mars. *Science* **259**, 1294–1297 (1993).
7. Jakosky, B. M., Henderson, B. G. & Mellon, M. T. Chaotic obliquity and the nature of the martian climate. *J. Geophys. Res.* **100**, 1579–1584 (1995).
8. Hartmann, W. K. *et al.* in *Basaltic Volcanism on the Terrestrial Planets* (eds Basaltic Volcanism Study Project) 1049–1127 (Pergamon, New York, 1981).
9. Tanaka, K. L. The stratigraphy of Mars. *J. Geophys. Res.* **91**, E139–E158 (1986).
10. Hartmann, W. K. & Berman, D. C. Elysium Planitia lava flows: crater count chronology and geological implications. *J. Geophys. Res.* **105**, 15011–15025 (2000).
11. Hartmann, W. K. & Neukum, G. Cratering chronology and the evolution of Mars. *Space Sci. Rev.* (in the press).
12. Pepin, R. O. Evolution of the martian atmosphere. *Icarus* **111**, 289–304 (1994).
13. Carr, M. H. & Clow, G. D. Martian channels and valleys: their characteristics, distribution, and age. *Icarus* **48**, 91–117 (1981).
14. Carr, M. H. & Malin, M. C. Meter-scale characteristics of martian channels and valleys. *Icarus* **146**, 366–386 (2000).
15. Carr, M. H. & Chuang, F. C. Martian drainage densities. *J. Geophys. Res.* **102**, 9145–9152 (1997).
16. Tanaka, K. L., Dohm, J. M., Lias, J. H. & Hare, T. M. Erosional valleys in the Thaumasia region of Mars: hydrothermal and seismic origins. *J. Geophys. Res.* **103**, 31407–31419 (1998).
17. Carr, M. H. *Water on Mars* (Oxford Univ. Press, New York, 1996).
18. Chapman, C. R. & Jones, K. L. Cratering and obliteration history of Mars. *Annu. Rev. Earth Planet. Sci.* **5**, 515–540 (1977).
19. Craddock, R. A. & Maxwell, T. A. Geomorphic evolution of the martian highlands through ancient fluvial processes. *J. Geophys. Res.* **98**, 3453–3468 (1993).
20. Craddock, R. A., Maxwell, T. A. & Howard, A. D. Crater morphometry and modification in the Sinus Sabaeus and Margaritifer Sinus regions of Mars. *J. Geophys. Res.* **102**, 13321–13340 (1997).
21. Golombek, M. P. & Bridges, N. T. Erosion rates on Mars and implications for climate change: constraints from the Pathfinder landing site. *J. Geophys. Res.* **105**, 1841–1853 (2000).
22. Baker, V. R. & Partridge, J. Small martian valleys: pristine and degraded morphology. *J. Geophys. Res.* **91**, 3561–3572 (1986).
23. Pollack, J. B., Kasting, J. F., Richardson, S. M. & Poliakov, K. The case for a warm, wet climate on early Mars. *Icarus* **71**, 203–224 (1987).
24. Squires, S. W. & Kasting, J. F. Early Mars: how warm and how wet? *Science* **265**, 744–749 (1994).
25. Kasting, J. F. CO₂ condensation and the climate of early Mars. *Icarus* **94**, 1–13 (1991).
26. Forget, F. & Pierrehumbert, R. T. Warming early Mars with carbon dioxide clouds that scatter infrared radiation. *Science* **278**, 1273–1276 (1997).
27. Mischna, M. A., Kasting, J. F., Pavlov, A. & Freedman, R. Influence of carbon dioxide clouds on early martian climate. *Icarus* **145**, 546–554 (2000).
28. Phillips, R. J. *et al.* Ancient geodynamics and global-scale hydrology on Mars. *Science* **291**, 2587–2591 (2001).
29. Anderson, R. C. *et al.* Primary centers and secondary concentrations of tectonic activity through time in the western hemisphere of Mars. *J. Geophys. Res.* (in the press).
30. McSweeney, H. J. Jr *et al.* Geochemical evidence for magmatic water within Mars from pyroxenes in the Shergottite meteorite. *Nature* **409**, 487–490 (2001).
31. Parker, T. J., Clifford, S. M. & Banerdt, W. B. Argyle Planitia and the Mars global hydrologic cycle. *Lunar Planet. Sci. Conf. XXXI*, Abstr. 2033 <<http://www.lpi.usra.edu/meetings/lpsc2000/pdf/2033.pdf>> (2000).
32. Baker, V. R. *The Channels of Mars* 198 p (Univ. Texas Press, Austin, TX, 1982).
33. Jakosky, B. M. & Phillips, R. J. Water the many mysteries of Mars? (Abstr.) Am. Geophys. Union Fall meeting, San Francisco <<http://www.agu.org/meetings/waisf00.html>> (2000).
34. Acuña, M. H. *et al.* Global distribution of crustal magnetism discovered by the Mars Global Surveyor MAG/ER experiment. *Science* **284**, 790–793 (1999).
35. Connerney, J. E. P. *et al.* Magnetic lineations in the ancient crust of Mars. *Science* **284**, 794–798 (1999).
36. Hynes, B. M. & Phillips, R. J. Evidence for extensive denudation of the Martian highlands. *Geology* **29**, 407–410 (2001).
37. Harrison, K. P. & Grimm, R. E. Martian hydrothermal systems: relationship between magnetic anomalies and valley networks. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1441 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1441.pdf>> (2001).
38. Newman, M. J. & Rood, R. T. Implications of solar evolution for the Earth's early atmosphere. *Science* **198**, 1035–1037 (1977).
39. Kasting, J. F. & Grinspoon, D. H. in *The Sun in Time* (eds Sonett, C. P., Giampapa, M. S. & Matthews, M. S.) 447–462 (Univ. Arizona Press, Tucson, 1991).
40. Haberle, R. M. Early Mars climate models. *J. Geophys. Res.* **103**, 28467–28479 (1998).
41. Melosh, H. J. & Vickery, A. M. Impact erosion of the primordial atmosphere of Mars. *Nature* **338**, 487–489 (1989).
42. Brain, D. A. & Jakosky, B. M. Atmospheric loss since the onset of the martian geologic record: combined role of impact erosion and sputtering. *J. Geophys. Res.* **103**, 22689–22694 (1998).
43. Frey, H. V., Shockey, K. M., Frey, E. L., Roark, J. H. & Sakimoto, S. E. H. A very large population of likely buried impact basins in the northern lowlands of Mars revealed by MOLA data. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1680 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1680.pdf>> (2001).
44. Chyba, C. F., Owen, T. C. & Ip, W. H. in *Hazards Due to Comets and Asteroids* (ed. Gehrels, T.) 9–58 (Univ. Arizona Press, Tucson, 1994).
45. Owen, T. & Bar-Nun, A. Comets, impacts, and atmospheres. *Icarus* **116**, 215–226 (1995).
46. Luhmann, J. G., Johnson, R. E. & Zhang, M. H. G. Evolutionary impact of sputtering of the martian atmosphere by O⁺ pickup ions. *Geophys. Res. Lett.* **19**, 2151–2154 (1992).
47. Mitchell, D. L. *et al.* Crustal magnetocylinders at Mars. (Abstr.) Am. Geophys. Union Spring meeting <<http://www.agu.org/meetings/waism00.html>> (2000).
48. Ayres, T. R. Evolution of the solar ionizing flux. *J. Geophys. Res.* **102**, 1641–1651 (1997).
49. McElroy, M. B. & Yung, Y. L. Oxygen isotopes in the martian atmosphere: implications for the evolution of volatiles. *Planet. Space Sci.* **24**, 1107–1113 (1976).
50. Jakosky, B. M., Pepin, R. O., Johnson, R. E. & Fox, J. L. Mars atmospheric loss and isotopic fractionation by solar-wind-induced sputtering and photochemical escape. *Icarus* **111**, 271–288 (1994).
51. Hutchins, K. S. & Jakosky, B. M. Evolution of martian atmospheric argon: implications for sources of volatiles. *J. Geophys. Res.* **101**, 14933–14949 (1996).
52. Owen, T., Maillard, J. P., deBergh, C. & Lutz, B. L. Deuterium on Mars: the abundance of HDO and the value of D/H. *Science* **240**, 1767–1770 (1988).
53. Bjoraker, G. L., Mumma, M. J. & Larson, H. P. Isotopic abundance ratios for hydrogen and oxygen in the martian atmosphere. *Bull. Am. Astron. Soc.* **21**, 991 (1989).
54. Krasnopolsky, V. A., Bjoraker, G. L., Mumma, M. J. & Jennings, D. E. High-resolution spectroscopy of Mars at 3.7 and 8 μ m: a sensitive search for H₂O, H₂CO, HCl, and CH₄, and detection of HDO. *J. Geophys. Res.* **102**, 6525–6534 (1997).
55. Liu, S. C. & Donahue, T. M. The regulation of hydrogen and oxygen escape from Mars. *Icarus* **28**, 231–246 (1976).
56. Yung, Y. L. *et al.* HDO in the martian atmosphere: implications for the abundance of crustal water. *Icarus* **76**, 146–159 (1988).
57. Leshin, L. A. Insights into martian water reservoirs from analyses of martian meteorite QUE94201. *Geophys. Res. Lett.* **27**, 2017–2020 (2000).
58. Krasnopolsky, V. On the deuterium abundance on Mars and some related problems. *Icarus* **148**, 597–602 (2000).
59. Jakosky, B. M. & Leshin, L. A. Mars D/H: implications for volatile evolution and climate history. (Abstr.) Am. Geophys. Union Spring meeting, Boston <<http://www.agu.org/meetings/waism01.html>> (2001).
60. Donahue, T. M. Evolution of water reservoirs on Mars from D/H ratios in the atmosphere and crust. *Nature* **374**, 432–434 (1995).
61. Gooding, J. L., Wentworth, S. J. & Zolensky, M. E. Calcium carbonate and sulfate of possible extraterrestrial origin in the EETA 79001 meteorite. *Geochim. Acta* **52**, 909–915 (1988).
62. Romanek, C. S. *et al.* Record of fluid-rock interactions on Mars from the meteorite ALH84001. *Nature* **372**, 655–657 (1994).
63. Treiman, A. H., Barrett, R. A. & Gooding, J. L. Preterrestrial alteration of the Lafayette (SNC) meteorite. *Meteoritics* **28**, 86–97 (1993).
64. Pollack, J. B. *et al.* Thermal emission spectra of Mars (5.4–10.5 μ m): evidence for sulfates, carbonates, and hydrates. *J. Geophys. Res.* **95**, 14595–14627 (1990).
65. Christensen, P. R. *et al.* Mars Global Surveyor Thermal Emission Spectrometer experiment: investigation, description and surface science results. *J. Geophys. Res.* (in the press).
66. Marti, K. & Mathew, K. J. Ancient martian nitrogen. *Geophys. Res. Lett.* **27**, 1463–1466 (2000).
67. Mathew, K. J. & Marti, K. Early evolution of martian volatiles: nitrogen and noble gas components in ALH84001 and Chassigny. *J. Geophys. Res.* **106**, 1401–1422 (2001).
68. Turner, G., Knott, S. F., Ash, R. D. & Gilmour, J. D. Ar–Ar chronology of the martian meteorite ALH84001: evidence for the timing of the early bombardment of Mars. *Geochim. Cosmochim. Acta* **61**, 3835–3850 (1997).
69. Hutchins, K. S., Jakosky, B. M. & Luhmann, J. G. Impact of a paleo-magnetic field on sputtering loss

- of martian atmospheric argon and neon. *J. Geophys. Res.* **102**, 9183–9189 (1997).
70. Schubert, G., Russell, C. T. & Moore, W. B. Timing of the martian dynamo. *Nature* **408**, 666–667 (2000).
71. Weiss, B. P. *et al.* Records of an ancient Martian magnetic field in ALH84001. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1244 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1244.pdf>> (2001).
72. Carr, M. H. Formation of martian flood features by release of water from confined aquifers. *J. Geophys. Res.* **84**, 2995–3007 (1979).
73. Hoffman, N. White Mars: a new model for Mars' surface and atmosphere based on CO₂. *Icarus* **146**, 326–342 (2000).
74. Lucchitta, B. K. Antarctic ice streams and outflow channels on Mars. *Geophys. Res. Lett.* **28**, 403–406 (2001).
75. Tanaka, K. L. Debris flow origin for the Simud/Tiu deposit on Mars. *J. Geophys. Res.* **104**, 8637–8652 (1999).
76. Carr, M. H. Mars: a water-rich planet? *Icarus* **68**, 187–216 (1986).
77. Watson, L. L., Hutcheon, I. D., Epstein, S. & Stolper, E. M. Water on Mars: clues from deuterium/hydrogen and water contents of hydrous phases in SNC meteorites. *Science* **265**, 86–90 (1994).
78. Jakosky, B. M. & Jones, J. H. The history of martian volatiles. *Rev. Geophys.* **35**, 1–16 (1997).
79. McSween, H. Y. Jr SNC Meteorites: clues to martian petrologic evolution? *Rev. Geophys.* **23**, 391–416 (1985).
80. McSween, H. Y. Jr What we have learned about Mars from SNC meteorites. *Meteoritics* **29**, 757–779 (1994).
81. Swindle, T. D. *et al.* Noble gases in iddingsite from the Lafayette meteorite: evidence for liquid water on Mars in the last few hundred million years. *Meteoritics Planet. Sci.* **35**, 107–115 (2000).
82. Christensen, P. R. *et al.* Detection of crystalline hematite mineralization on Mars by the Thermal Emission Spectrometer: evidence for near-surface water. *J. Geophys. Res.* **105**, 9623–9642 (2000).
83. Tanaka, K., Chapman, M., Johnson, J. & Titus, T. Examination of igneous alternatives to Martian hematite using terrestrial analogs. *GSA Abstr. Programs* **32**(7), Abstr. 52142 (2000).
84. Malin, M. C. & Edgett, K. S. Evidence for recent ground water seepage and surface runoff on Mars. *Science* **288**, 2330–2335 (2000).
85. Musselwhite, D. S., Swindle, T. D. & Lunine, J. I. Liquid CO₂ breakout and the formation of recent small gullies on Mars. *Geophys. Res. Lett.* **28**, 1283–1285 (2001).
86. Stewart, S. T. & Nimmo, F. Surface runoff features on Mars: testing the carbon dioxide formation hypothesis. *J. Geophys. Res.* (submitted).
87. Mellon, M. T. & Phillips, R. J. Recent gullies on Mars and the source of liquid water. *J. Geophys. Res.* (in the press).
88. Cabrol, N. A. & Grin, E. A. Distribution, classification, and ages of martian impact crater lakes. *Icarus* **142**, 160–172 (1999).
89. Malin, M. C. & Edgett, K. S. Sedimentary rocks of early Mars. *Science* **290**, 1927–1937 (2001).
90. Parker, T. S., Saunders, R. S. & Schneeberger, D. M. Transitional morphology in the west Deuteronilus Mensae region of Mars: implications for modification of the lowland/upland boundary. *Icarus* **82**, 111–145 (1989).
91. Parker, T. J., Gorsline, D. S., Saunders, R. S., Pieri, D. & Schneeberger, D. M. Coastal geomorphology of the martian northern plains. *J. Geophys. Res.* **98**, 11061–11078 (1993).
92. Head, J. W. *et al.* Possible ancient oceans on Mars: evidence from Mars Orbiter Laser Altimeter. *Science* **286**, 2134–2137 (1999).
93. Head, J. W. III *et al.* Oceans in the past history of Mars: tests for their presence using Mars Orbiter Laser Altimeter (MOLA) data. *Geophys. Res. Lett.* **25**, 4401–4404 (1998).
94. Malin, M. C. & Edgett, K. S. Oceans or seas in the martian northern lowlands: high-resolution imaging tests of proposed coastlines. *Geophys. Res. Lett.* **26**, 3049–3052 (1999).
95. Withers, P. & Neumann, G. A. Enigmatic northern plains of Mars. *Nature* **410**, 651 (2001).
96. Scott, D. H. & Tanaka, K. L. Geologic map of the western equatorial region of Mars. US Geol. Surv. Map 1-1802-A (1986).
97. Aharonson, O., Zuber, M. T., Neumann, G. A. & Head, J. W. Mars: northern hemisphere slopes and slope distributions. *Geophys. Res. Lett.* **25**, 4413–4416 (1998).
98. Smith, D. E. *et al.* The global topography of Mars and implications for surface evolution. *Science* **284**, 1495–1503 (1999).
99. Head, J. W., Kreslavsky, M. A. & Pratt, S. Northern lowlands on Mars: evidence for widespread volcanic flooding and tectonic deformation in the Early Hesperian. *Lunar Planet. Sci. Conf. XXXII*, Abstr. 1063 <<http://www.lpi.usra.edu/meetings/lpsc2001/pdf/1063.pdf>> (2001).
100. Komar, P. D. Modes of sediment transport in channelized water flows with ramifications to the erosion of the martian outflow channels. *Icarus* **42**, 317–329 (1980).
101. Zuber, M. T. *et al.* Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* **287**, 1788–1793 (2000).
102. Pechmann, J. C. The origin of polygonal troughs on the northern plains of Mars. *Icarus* **42**, 185–210 (1980).
103. Hiesinger, H. & Head, J. W. Characteristics and origin of polygonal terrain in southern Utopia Planitia, Mars: results from Mars Orbiter Laser Altimeter and Mars Orbiter Camera data. *J. Geophys. Res.* **105**, 11999–12022 (2000).
104. Lane, M. D. & Christensen, P. R. Convection in a catastrophic flood deposit as the mechanism for the giant polygons on Mars. *J. Geophys. Res.* **105**, 17617–17627 (2000).
105. Greeley, R. & The Mars Exploration Payload Advisory Group. *Mars Exploration Program: Scientific Goals, Objectives, Investigations, and Priorities* (Jet Propulsion Laboratory Publication, in the press).
106. Purucker, M. *et al.* An altitude-normalized magnetic map of Mars and its interpretation. *Geophys. Res. Lett.* **27**, 2449–2452 (2000).
107. Goldspiel, J. M. & Squyres, S. W. Ancient aqueous sedimentation on Mars. *Icarus* **89**, 392–410 (1991).

Acknowledgements

We thank C. Leovy, S. Stewart, L. Leshin, M. Mellon, H. Frey, P. Withers, B. Hynek, K. Harrison, W. Hartmann and the MOLA science team for valuable discussions and for providing preprints of their manuscripts. We also thank J. Head, R. Haberle and C. Leovy for detailed reviews of our manuscript. This research was supported by the Mars Global Surveyor Project and the NASA Planetary Geology and Geophysics Program.

Weather and climate on Mars

Conway Leovy

Department of Atmospheric Sciences, Box 351640, University of Washington, Seattle, Washington, 98195, USA
(e-mail: conway@atmos.washington.edu)

Imagine a planet very much like the Earth, with similar size, rotation rate and inclination of rotation axis, possessing an atmosphere and a solid surface, but lacking oceans and dense clouds of liquid water. We might expect such a desert planet to be dominated by large variations in day–night and winter–summer weather. Dust storms would be common. Observations and simulations of martian climate confirm these expectations and provide a wealth of detail that can help resolve problems of climate evolution.

Aside from the very limited atmospheric water and associated latent heat release, Mars' atmosphere differs from Earth's primarily in its very low surface pressure, which varies as much as 20% annually owing to seasonal CO₂ condensation at the poles^{1,2} (Table 1). The corresponding low atmospheric heat capacity and short bulk radiative timescale, $\langle \tau_r \rangle$, ensure that diurnal and seasonal temperature variations are much larger than those in terrestrial deserts. Maximum diurnal near-surface atmospheric temperature range is about 30 °C on Earth and about 60 °C on Mars. Seasonal daily averages of surface air temperatures range up to 90 °C in martian polar regions¹, but only up to about 50 °C in the extreme terrestrial climate of northeastern Siberia.

Martian atmospheric temperature is strongly controlled by suspended dust^{3,4}. When the atmosphere is 'clear' (optical depth at visible wavelengths $\ll 1$), daytime surface heating generates convection that extends as high as 1 scale height (about 10 km, see Table 1). In this convective boundary layer, temperature decreases with height at the dry adiabatic lapse rate, about 4.3 K km⁻¹. Above this layer, temperature declines more gradually with height and can approach the limiting condensation temperature of CO₂ (130–150 K) above 40 km (refs 5, 6; Fig. 1). During the night, convection collapses and a strong temperature inversion develops in the lowest kilometre. Above 40 km, there are strong time-dependent oscillations of the vertical temperature profile, probably produced by vertically propagating waves⁶. Near winter solstice, Mars' distance from the Sun approaches its minimum, and the atmosphere becomes dusty (optical depth at visible wavelengths ≥ 1). Absorption of solar radiation by dust increases the daily mean temperature and diurnal temperature range above the surface, but the surface diurnal temperature range decreases and the daytime convective layer disappears. Daily average temperature near the surface is not affected by dust loading.

Horizontal variations of temperature and associated horizontal variations of pressure drive the planetary wind system — the general circulation. Our knowledge of the general circulation of Mars is based on observations and numerical simulation models. Until very recently, these models were sophisticated but poorly constrained by data. Recent observations from the Mars Global Surveyor (MGS) now provide a good match between the degree of detail produced by the models and that available from observations. The following sections describe our present understanding of the general circulation and focus on the critical interaction between surface winds and dust storms in the present climate regime. The last section discusses the research challenge — connecting

our understanding of the present climate to an improved understanding of the evolution of martian climate and surface features over the past four billion years.

The general circulation

Wind measurements are limited to surface winds at lander sites, vertical profiles of wind from the two Viking lander entry vehicles, and scattered inferences of wind from cloud features. Vertical temperature profiles inferred from thermal emission measurements, obtained primarily by Mars orbiting spacecraft, provide far more complete coverage and can be used to infer winds^{7–9}. The MGS Thermal Emission Spectrometer (MGS-TES) has provided vertical temperature profiles that resolve features whose vertical scale is about 1 scale height or more⁹. In addition, Earth-based microwave measurements provide information about atmospheric temperature distribution¹⁰. Atmospheric

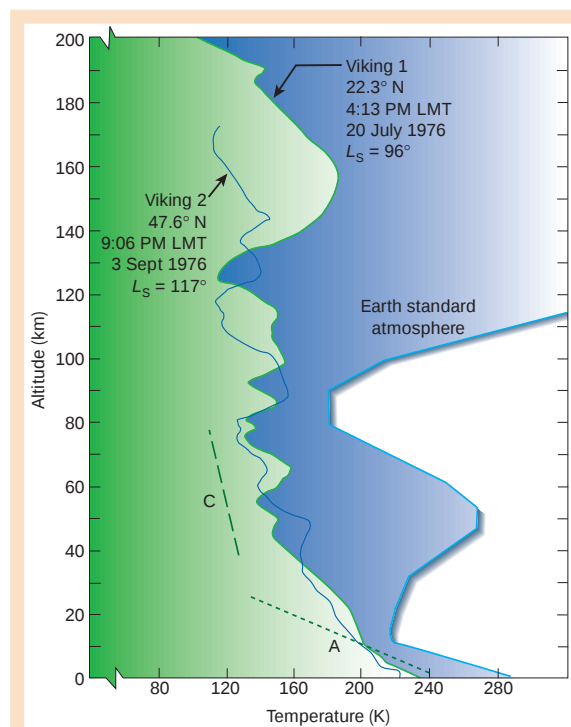


Figure 1 Mars atmosphere temperature profiles measured by Viking⁵ compared with the standard atmosphere profile for Earth. Martian adiabatic lapse rate and CO₂ condensation are indicated by lines A and C, respectively. The Viking profiles are reproduced by permission of the American Geophysical Union.

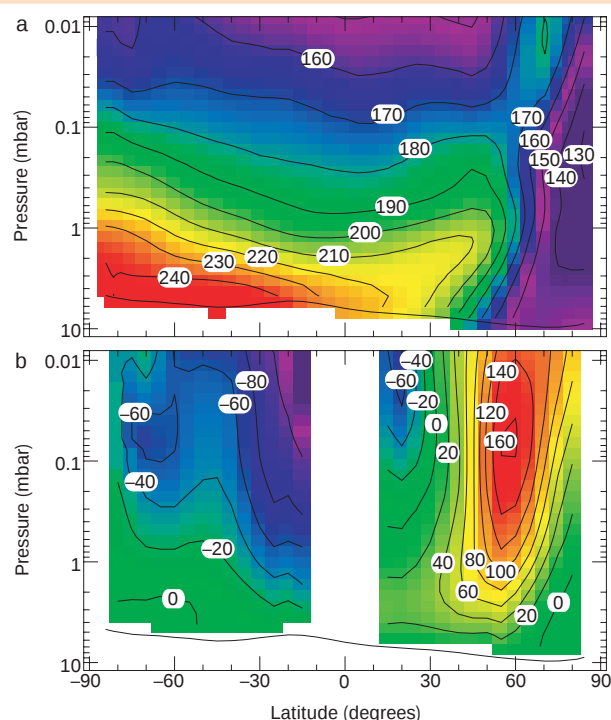


Figure 2 Cross-sections of zonal mean temperature and zonal wind for northern-hemisphere winter solstice ($L_s = 270^\circ$) from MGS-TES observations. **a**, Temperature (K); **b**, eastward wind (m s^{-1}). Wind was derived from the thermal gradient wind approximation under the assumption that surface wind is zero everywhere (Box 1). The ordinate scale of log pressure corresponds approximately to height, with each decade roughly equivalent to 20 km, and the upper boundary near 60 km. Figure adapted from ref. 9; reproduced by permission of the American Geophysical Union.

refractive index inferred from Doppler delay of radio signals passing through the atmosphere (radio occultation) provides a limited number of temperature profiles with much higher vertical resolution¹¹. These also give information about the distribution of pressure on horizontal surfaces. Time series of surface pressure are available from the sites of two Viking and one Pathfinder landed spacecraft^{6,12}. Three-dimensional time-dependent distributions of temperature together with pressure on a single horizontal surface can be used to derive approximate large-scale wind fields (see Box 1). Wind distributions can also be derived through the process of assimilation of temperature data into more complete nonlinear dynamical models¹³.

A representative MGS-TES temperature cross-section in the meridional plane and the corresponding zonal mean (latitude average) eastward wind field derived from thermal gradient wind balance is shown in Fig. 2. It can be compared with a meridional cross-section of temperature and total (not approximate) zonal mean eastward wind generated by a martian numerical general circulation model (GCM; Fig. 3a, b). Agreement between observed temperatures and the derived thermal gradient winds and the temperatures and winds generated by the GCM is remarkable and serves as a good test of our ability to simulate the martian general circulation^{14–16}. Viking landing site and radio occultation measurements also provide some validation for the GCM surface winds¹¹. Zonal mean mid-latitude eastward winds increase with height to maxima (jet streams) near or above the level of 3 scale heights (about 30 km) in all seasons except mid-summer, with westward wind throughout most of the subtropics and tropics. Similar zonal mean wind distributions are found on Earth, but they are weaker and the mean eastward jet streams occur at lower altitudes and latitudes. Eastward zonal winds prevail during summer at mid-latitudes on Earth, but mid-latitude zonal winds are generally westward on Mars

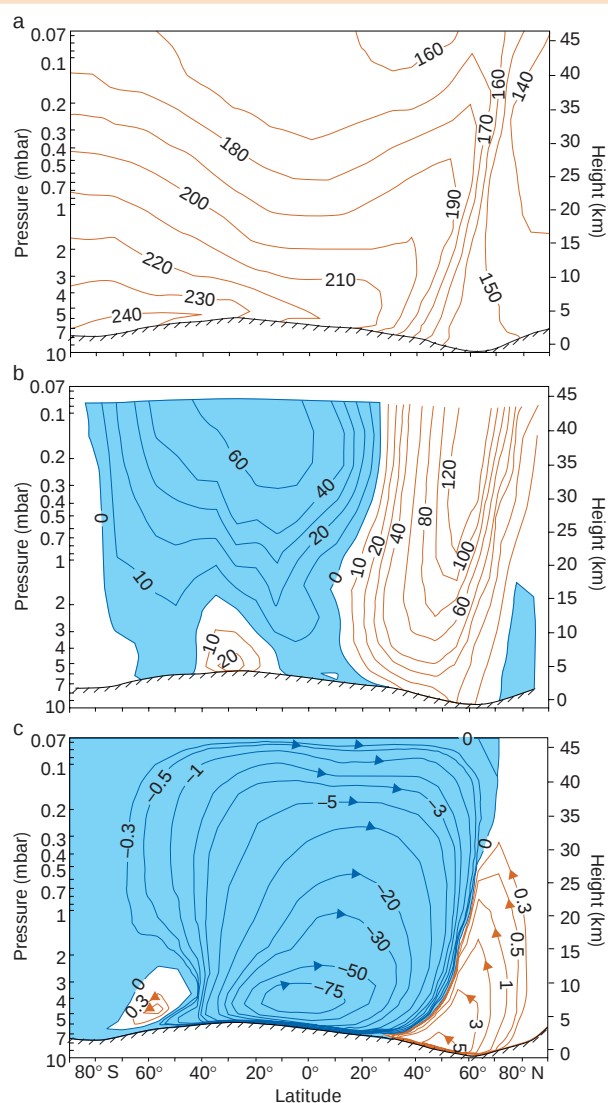


Figure 3 Zonal mean general circulation components from the GCM of Haberle et al.¹⁴. **a**, Temperature (K); **b**, west-east wind (m s^{-1} ; shading corresponds to westward wind); and **c**, meridional mass stream function (10^8 kg s^{-1} ; shading corresponds to clockwise circulation in the plane of the page). Simulation is for the northern winter solstice with average optical depth at visible wavelengths equal to 0.3, slightly lower than observed at this season by MGS-TES. Reproduced by permission of the American Geophysical Union.

in mid-summer. These differences in zonal winds and associated meridional temperature gradients between Mars and Earth were predicted long before spacecraft measurements became available¹⁷. They can be understood in terms of the thermal response to underlying oceans on Earth and underlying desert on Mars, together with the much shorter martian radiative timescale (Table 1).

The zonal mean meridional circulation, consisting of the zonally averaged poleward and upward flow components, can also be deduced approximately from the temperature field¹⁸ (see Box 2). On Earth, this meridional circulation consists of a tropical circulation cell and a weaker mid-latitude circulation cell. The tropical cell, called the Hadley circulation, is often described in terms of an ascending branch on the Equator and descending branches near 30° in the subtropics of both hemispheres. In fact, this symmetry about the Equator is observed only near equinox. In northern winter, Earth's Hadley circulation is dominated by a single cell with ascending motion south of the Equator and descending motion in the

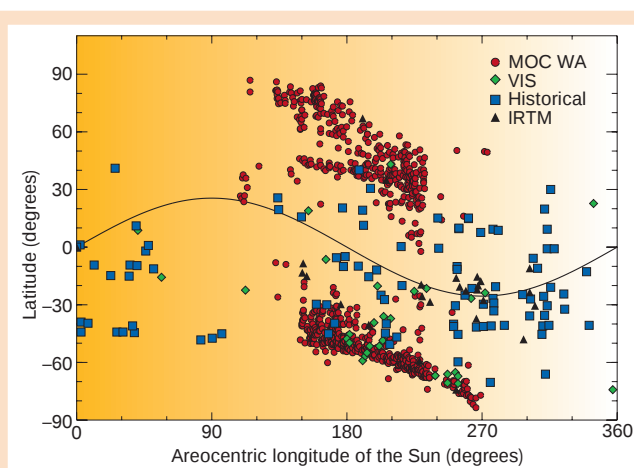


Figure 4 Local-to-regional dust storms observed by Cantor *et al.*⁴³. The horizontal axis L_s is Mars-centred longitude of the Sun, an index of season. $L_s = 0^\circ$ corresponds to northern spring equinox; $L_s = 270^\circ$ corresponds to northern winter solstice. The figure shows all storms in the size range 10^2 – 1.6×10^6 km² observed between $L_s = 107.27$ and $L_s = 262.78$ in MGS-MOC wide angle (WA) images. Viking orbiter and historical ground-based observations of dust storms are also shown⁴. VIS and IRTM correspond respectively to storms observed by the Viking Orbiter Visible Imager and Infrared Thermal Mapper. Reproduced by permission of the American Geophysical Union.

northern subtropics. This pattern reverses during southern winter¹⁹. Similar asymmetric Hadley circulation prevails on Mars, but the circulation cells are much stronger and the ascending and descending branches are displaced much farther from the equator (30–60° latitudes)^{14,20} (Fig. 3c). A symmetric, but very weak equatorially symmetric, Hadley circulation does occur on Mars, but only near the equinoxes. Because of Mars' high orbital eccentricity and resulting large seasonal differences in heating, its Hadley circulation is stronger and wider during northern winter than southern winter. Bright depositional streaks on the surface reflect the dominant Hadley circulation pattern of surface winds during this season^{21,22}. In the temperature distribution of Fig. 2a, the descending branch of the Hadley circulation is evident in the temperature maximum in the latitude belt 50–60° N of the winter hemisphere.

The symmetry of the zonal mean west–east wind pattern is broken in middle latitudes by prominent large-scale wave-like flow disturbances — planetary waves. These waves are observationally well constrained from measurements at the two Viking lander sites and from MGS-TES radiance measurements. Travelling planetary waves with zonal wave numbers (number of waves around a latitude circle) from 1 to 3 occur during winter and spring in mid- and high latitudes of both hemispheres^{23–25}. These propagate eastward, are well organized, and are apparently lower wave-number counterparts of eastward propagating mid-latitude weather systems on Earth. They arise from instability of the zonal mean temperature and wind fields (baroclinic and/or barotropic instability). These systems were predicted in advance of any observations¹⁷, and they are well simulated by GCMs. Other weather systems of intermediate scale (with diameters of a few hundred kilometres) occur near the edges of the residual mid-summer north polar cap^{26,27}. These systems, which produce familiar-looking moving comma-shaped cloud patterns, seem to be dry counterparts of Earth's polar lows that form primarily over open water near the sea-ice edge during winter. Stationary planetary waves (mainly zonal wave-numbers 1 and 2), generated by interaction between eastward zonal winds and the large-amplitude topography, are also observed and simulated in GCMs^{28,29}. In addition, flow over topography generates internal gravity waves (or 'buoyancy waves') whose horizontal scales range from a few kilometres to several hundred kilometres (ref. 30).

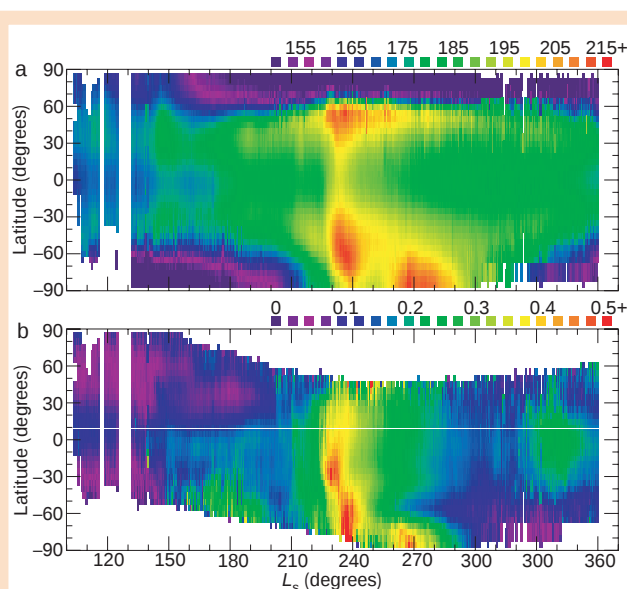


Figure 5 Temperature and dust opacity versus latitude and seasonal index (L_s) as inferred from MGS-TES by Smith *et al.*⁹. **a**, Temperature is at the 0.5-hPa level (approximately 20 km), and **b**, dust opacity at spectral wave-number 1,075 cm^{−1} is at the surface. Note the close correspondence between increased southern-hemisphere dust opacity and increased northern-hemisphere temperature, beginning near $L_s = 230$. This illustrates the global response of the Hadley circulation to enhanced heating in the southern (summer) hemisphere. Reproduced by permission of the American Geophysical Union.

The strong diurnal heating cycle generates a family of planetary waves known as thermal tides^{15,31–33}; these are centred in the tropics, but extend to mid-latitudes. Evidence for thermal tides derives from large-amplitude fluctuations in vertical temperature profiles above about 40 km (Fig. 1). Most thermal tides follow the Sun westward, but interactions between Sun-following tidal components and the large-amplitude planetary topography generate additional waves. The most prominent of these are equatorially centred, eastward-propagating Kelvin waves^{15,34}. Some of the energy of thermal tides and Kelvin waves propagates to elevations greater than 100 km in the upper atmosphere³⁵. Absorption of upward-propagating tides and the smaller-scale internal gravity waves can strongly influence the background flow at altitudes above about 40 km (ref. 36). These waves tend to drive the background flow towards the horizontal phase speed of the waves.

Surface winds and dust storms

Surface winds calculated by GCMs are generally consistent with the few available direct measurements³⁷ and with the near-surface branch of the Hadley circulation inferred from MGS-TES temperature and surface dust streaks^{21,22}. As on Earth, the Hadley circulation is associated with 'trade winds' blowing from the northeast in the northern hemisphere and the southeast in the southern hemisphere. These winds are confined below $\pm 30^\circ$ latitude on Earth, but extend as high as 50° latitude during the martian northern winter. During the solstice seasons, martian trade winds cross the equator, curve towards the east and form a low-elevation eastward jet in the summer subtropics¹¹ (Fig. 3b). This picture describes the zonal mean, but, as on Earth, the Hadley circulation winds are not longitudinally uniform. The strongest low-latitude martian surface wind probably occurs in lowlands and/or on east-facing slopes of large-scale topographic rises³⁸.

Strong surface wind can set particles of ~ 100 - μ m diameter into saltating motion. The resulting surface impacts can raise smaller particles into suspension to produce dust storms. Threshold wind

Box 1 Thermal wind balance

If the planetary Rossby number $Ro \ll 1$ (see Table 1), the horizontal wind vector $\mathbf{u}$ and pressure p are related by the geostrophic wind approximation: $\mathbf{u} = gH(pf)^{-1} \mathbf{k} \nabla_h p$, where $f = 2\Omega \sin \phi$ is the Coriolis parameter at latitude ϕ , ∇_h is the gradient operator on a horizontal surface, and $\mathbf{k}$ is a 90° counterclockwise rotation operator. With the hydrostatic balance equation and the ideal gas law, this equation implies the thermal wind equation relating the vertical shear of $\mathbf{u}$ to temperature T : $\partial(\mathbf{u}/T)/\partial z = g(fT^2)^{-1} \mathbf{k} \nabla_h T$. These approximations are usually accurate, with errors of less than 10%, for large-scale extra-tropical winds on Earth, but they must be applied with caution to Mars because Ro is larger there (Table 1). However, a slightly modified version of the thermal wind equation, the zonal thermal gradient wind equation, is very accurate outside of the deep tropics for zonal mean eastward wind u on Mars as well as Earth:

$\partial(u/T)/\partial z = -g(f^*T^2)^{-1} \partial T/\partial y$, where $f^* = (f + u \tan \phi/a)$. To apply this equation, knowledge of u on some horizontal surface is required. Wind near the ground is generally much weaker than wind aloft, and surface wind can be set equal to zero as a first approximation (ref. 1, pp. 854–856).

speed for initiation of dust storms depends on surface roughness, size distribution of fine particles, and surface pressure and temperature³⁹. At the two Viking lander sites, this threshold is apparently about 30 m s⁻¹ for winds at 2-m height^{40,41}. Although, at most sites, such strong wind speeds are rare, over an annual cycle and the planet as a whole, many local dust storms occur^{42,43}. Daytime convection also generates dust devils that can extend through the depth of the convective boundary layer^{44,45}.

As seen in MGS Mars Orbital Camera (MGS-MOC) images, there is a strong tendency for dust storms in the size range 10⁴–10⁶ km² to occur in mid-latitude storm zones following the edge of the seasonally varying polar caps (Fig. 4). Dust storms also occur in lower latitudes, especially during northern winter when Mars is closest to the Sun. These low-latitude storms seem to be generated by a combination of the strong seasonal Hadley circulation, thermal tides, topographically forced flow and local thermal convection⁴⁶. Dust

Box 2 Zonal mean meridional circulation

In the meridional (vertical and north–south) plane, the vertical component of the flow, w , follows from an approximate version of the thermodynamic equation: $w(\Gamma_a - \Gamma) = Q/C_p = -\tau_r^{-1}(T - T_e)$, where Q is net diabatic heating rate per unit mass, T_e is local radiative equilibrium temperature, τ_r is local radiative adjustment timescale, and Γ_a and Γ are the dry adiabatic and actual lapse rates (see Table 1, and ref. 1, p. 873). Because, in the zonal average, $\Gamma_a > \Gamma$, vertical velocity is downward where $T > T_e$, and vice versa. Both T_e and τ_r depend on the dust distribution as well as on latitude, height and season, but the dust distribution can also be inferred from orbiter data⁹. The poleward wind component v can be derived from w using the zonal average continuity equation $(a \cos \phi)^{-1} \partial(v \cos \phi) = \rho^{-1} \partial(\rho w)/\partial z$.

from low-latitude storms is carried aloft by the ascending branch of the Hadley circulation and spreads widely, so that the average optical depth of the atmosphere increases from a few tenths to ≥ 1 near northern winter solstice. Suspended dust strongly heats the atmosphere, inducing a powerful feedback: seasonal low-latitude winds lift dust, which leads to dust heating, which causes enhanced and enlarged Hadley circulation and thermal tides, which results in enhanced low-latitude winds. Enhancement and enlargement of the Hadley circulation is observed as dust is pumped high into the atmosphere^{47–49}. This appears as an intensification and slight poleward shift of the temperature maximum in the winter-hemisphere descending branch of the Hadley circulation that coincides closely with dust-storm events in the summer hemisphere (Fig. 5). The evolution of complex dust-storm events like this one and the corresponding atmospheric dynamical response are well simulated by GCMs^{15,50}. During some years, winter-solstice subtropical dust storms can expand to encircle the planet or even cover the globe⁵¹.

Short radiative timescale and strong control of temperatures by orbital and surface properties tend to lock the martian general circulation into predictably repetitive diurnal and annual cycles, but the occasional occurrence of planet encircling and global dust storms breaks this pattern and remains poorly understood⁵².

The research challenge

Lifting and transport of dust produce non-uniform surface erosion and deposition⁵³ and have substantially modified the surface of Mars over the past four billion years. If periods of higher surface pressure occurred in the past, threshold wind speeds for saltation would have been lower, and wind would have been a much more effective agent of surface modification. Variations in orbital parameters should also have produced substantial changes in the magnitude and distribution of net dust transport. Variations in tilt of the rotation axis (obliquity), whose range is approximately 15–40° in the present epoch, eccentricity, and argument of perihelion will generate large changes in the intensity and size of the Hadley circulation and in planetary waves. Because these meteorological factors interact with the large-amplitude martian topography, which is highly asymmetric about the equator, substantial transport between low- and high-elevation regions as well as substantial meridional transport should have occurred over the past four billion years. These factors complicate the geological interpretation of surface features. One of the main challenges is to integrate martian GCM simulations of past climates with studies of surface geology to decipher the record of martian climate history and the roles of impacts, volcanism, wind, ice and water in modifying the surface over time (see review in this issue by Jakosky and Phillips, pages 237–244). The good agreement between observations, theory and models of the present martian climate provides a solid basis for addressing this challenge. □

Table 1 Parameters governing the general circulations of Earth and Mars

	Earth	Mars
Mean distance from the Sun (AU)	1.0	1.52
Orbital eccentricity	0.03	0.093
Inclination of rotation axis to ecliptic	23.5°	25°
Rotation rate, Ω (10 ⁻⁴ s ⁻¹)	0.729	0.709
Planetary radius, a (km)	6,380	3,390
Surface gravity, g (m s ⁻²)	9.81	3.72
Representative surface pressure*, p (hPa)	1,000	7 ± 1
Bulk atmospheric composition (mole fraction)	N ₂ (0.8), O ₂ (0.2), A(0.001), CO ₂ (0.00037), H ₂ O(<0.03)	CO ₂ (0.95), N ₂ (0.03), A(0.02), H ₂ O(<0.0005)
Representative temperature of lowest scale height (K)	260	200
Specific gas constant, R (m ² s ⁻² K ⁻¹)	287	192
Representative scale height, RT/g (km)	7.6	10.3
Specific heat at constant pressure, C_p (m ² s ⁻² K ⁻¹)	1,000	860
Dry adiabatic lapse rate, $\Gamma_a (= -g/C_p)$ (K km ⁻¹)	9.8	4.3
Average lapse rate of lowest scale height, Γ (K km ⁻¹)	6.5	2.5
Bulk radiative timescale, $\langle \tau_r \rangle$ (10 ⁵ s)	40	2
Typical zonal wind at jet level, U (m s ⁻¹)	30	80
Planetary Rossby number, $Ro (= U/\Omega a)$	0.05	0.2

*Mars surface-pressure data correspond to approximate global mean annual average and annual range, not uncertainty.

1. Zurek, R. et al. in Mars (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 835–933 (Univ. Arizona Press, Tucson, 1992).
2. James, P., Kieffer, H. & Paige, D. in Mars (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 934–968 (Univ. Arizona Press, Tucson, 1992).
3. Gierasch, P. & Goody, R. The effect of dust on the temperature of the Mars atmosphere. *J. Atmos. Sci.* **29**, 400–402 (1972).
4. Kahn, R., Martin, T., Zurek, R. & Lee, S. in Mars (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 1017–1053 (Univ. Arizona Press, Tucson, 1992).
5. Seiff, A. & Kirk, D. Structure of the atmosphere of Mars in summer at mid-latitudes. *J. Geophys. Res.* **82**, 4364–4378 (1977).
6. Schofield, J. et al. The Mars Pathfinder Atmospheric Structure Investigation/Meteorology (ASI/MET) experiment. *Science* **278**, 1752–1758 (1997).
7. Martin, T. & Kieffer, H. Thermal infrared measurements of the martian atmosphere 2. The 15 μm band measurements. *J. Geophys. Res.* **84**, 2843–2852 (1979).
8. Wilson, R. & Richardson, M. The martian atmosphere during the Viking mission: infrared measurements of atmospheric temperatures revisited. *Icarus* **145**, 555–579 (2000).
9. Smith, M. et al. TES observations of atmospheric thermal structure and aerosol distribution during MGS mapping. *J. Geophys. Res.* **106** (in the press).
10. Clancy, R. et al. An intercomparison of ground-based millimeter, MGS TES, and Viking atmospheric temperature measurements: seasonal and interannual variability of temperatures and dust loading in the global Mars atmosphere. *J. Geophys. Res.* **105**, 9553–9571 (2000).
11. Hinson, D. et al. Initial results from radio occultation measurements with Mars Global Surveyor. *J. Geophys. Res.* **104**, 26997–27012 (1999).
12. Tillman, J., Johnson, N., Guttrop, P. & Percival, D. The martian annual pressure cycle: years without great dust storms. *J. Geophys. Res.* **98**, 10963–10971 (1993).
13. Lewis, S. et al. A climate database for Mars. *J. Geophys. Res.* **104**, 24177–24194 (1999).
14. Haberle, R. et al. Mars atmospheric dynamics as simulated by the NASA/Ames general circulation model. 1. The zonal-mean circulation. *J. Geophys. Res.* **102**, 13301–13311 (1993).
15. Wilson, J. & Hamilton, K. Comprehensive model simulation of thermal tides in the martian atmosphere. *J. Atmos. Sci.* **53**, 1290–1326 (1996).
16. Forget, F. et al. Improved general circulation models of the Martian atmosphere from the surface to above 80 km. *J. Geophys. Res.* **104**, 24156–24175 (1999).
17. Mintz, Y. in *The Atmospheres of Mars and Venus* (eds Kellogg, W. & Sagan, C.) NAS-NRC Publication 944, 107–146 (National Research Council, Washington DC, 1961).
18. Santee, M. & Crisp, D. Thermal structure and dust loading of the martian atmosphere during late summer: Mariner 9 revisited. *J. Geophys. Res.* **98**, 3261–3279 (1993).
19. Hartmann, D. *Global Physical Climatology* 140–143 (Academic, San Diego, 1994).
20. Conrath, B. et al. Mars Global Surveyor Thermal Emission Spectrometer (TES) observations: atmospheric temperatures during aerobraking and science phasing. *J. Geophys. Res.* **104**, 9509–9519 (1999).
21. Thomas, P., Veverka, J., Gineris, D. & Wong, L. “Dust” streaks on Mars. *Icarus* **49**, 398–415 (1984).
22. Greeley, R., Skyeck, A. & Pollack, J. Martian aeolian features and deposits: comparisons with general circulation model results. *J. Geophys. Res.* **98**, 3183–3196 (1993).
23. Barnes, J. Midlatitude disturbances in the Martian atmosphere: a second Mars year. *J. Atmos. Sci.* **38**, 225–234 (1981).
24. Barnes, J. Linear baroclinic instability in the Martian atmosphere. *J. Atmos. Sci.* **41**, 1536–1550 (1984).
25. Hollingsworth, J. et al. Orographic control of storm zones on Mars. *Nature* **380**, 413–416 (1996).
26. Gierasch, P., Thomas, P., French, R. & Veverka, J. Spiral clouds on Mars: a new atmospheric phenomenon. *Geophys. Res. Lett.* **6**, 405–408 (1979).
27. James, P., Hollingsworth, J., Wolff, J. & Lee, S. North polar dust storms in early spring on Mars. *Icarus* **38**, 64–73 (1999).
28. Conrath, B. Planetary-wave structure in the Martian atmosphere. *Icarus* **48**, 246–255 (1981).
29. Hollingsworth, J. & Barnes, J. Forced stationary planetary waves in Mars’s winter atmosphere. *J. Atmos. Sci.* **53**, 428–448 (1996).
30. Briggs, G. & Leovy, C. Mariner 9 observations of the Mars north polar hood. *Bull. Am. Meteorol. Soc.* **55**, 278–296 (1972).
31. Zurek, R. Diurnal tide in the martian atmosphere. *J. Atmos. Sci.* **33**, 321–337 (1976).
32. Zurek, R. & Leovy, C. Thermal tides in the dusty Martian atmosphere: a verification of theory. *Science* **213**, 437–439 (1981).
33. Hinson, D., Hollingsworth, J., Wilson, R. & Tyler, G. Radio occultation measurements of forced atmospheric waves on Mars. *J. Geophys. Res.* **106** (in the press).
34. Tillman, J. Mars global atmospheric oscillations: annually synchronized, transient normal mode oscillations and the triggering of global dust storms. *J. Geophys. Res.* **93**, 9433–9451 (1988).
35. Keating, G. et al. Evidence for large global diurnal Kelvin wave in the Mars upper atmosphere. *Bull. Am. Astron. Soc.* **32**, Abstr. 50:02 (2000).
36. Zurek, R. & Haberle, R. Zonally symmetric response to atmospheric tidal forcing in the dusty Martian atmosphere. *J. Atmos. Sci.* **45**, 2469–2485 (1988).
37. Murphy, J., Leovy, C. & Tillman, J. Observations of martian surface winds at the Viking Lander 1 site. *J. Geophys. Res.* **95**, 14555–14576 (1990).
38. Joshi, M., Lewis, S., Read, P. & Catling, D. Western boundary currents in the atmosphere of Mars. *Nature* **367**, 548–551 (1994).
39. Greeley, R., Lancaster, N., Lee, S. & Thomas, P. in Mars (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 835–933 (Univ. Arizona Press, Tucson, 1992).
40. Ryan, J., Sharman, R. & Lucich, R. Local Mars dust storm generation mechanism. *Geophys. Res. Lett.* **8**, 899–901 (1981).
41. Arvidson, R. et al. Three Mars years: Viking lander imaging observations. *Science* **222**, 463–468 (1983).
42. Peterfreund, A. & Kieffer, H. Thermal and infrared properties of the martian atmosphere. 3: Local dust clouds. *J. Geophys. Res.* **84**, 2853–2862 (1979).
43. Cantor, B., James, P., Caplinger, M. & Wolff, M. Martian dust storms: 1999 Mars Orbiter Camera observations. *J. Geophys. Res.* **106** (in the press).
44. Ryan, J. & Carroll, J. Dust devil wind velocities: mature state. *J. Geophys. Res.* **75**, 531–541 (1970).
45. Thomas, P. & Gierasch, P. Dust devils on Mars. *Science* **230**, 175–177 (1985).
46. Leovy, C., Zurek, R. & Pollack, J. Mechanisms of Mars dust storms. *J. Atmos. Sci.* **30**, 749–762 (1973).
47. Leovy, C., Tillman, J., Guest, W. & Barnes, J. in *Recent Advances in Planetary Meteorology* (ed G. Hunt) 69–84 (Cambridge Univ. Press, Cambridge, 1985).
48. Anderson, E. & Leovy, C. Mariner 9 television limb observations of dust and ice hazes on Mars. *J. Atmos. Sci.* **35**, 723–734 (1978).
49. Smith, M., Pearl, J., Conrath, B. & Christensen, P. Mars Global Surveyor Thermal Emission Spectrometer (TES) observations of dust opacity during aerobraking and science phasing. *J. Geophys. Res.* **105**, 9539–9552 (2000).
50. Murphy, J. et al. Three-dimensional numerical simulation of Martian global dust storms. *J. Geophys. Res.* **100**, 26357–26376 (1995).
51. Zurek, R. & Martin, L. Interannual variability of planet-encircling dust storms on Mars. *J. Geophys. Res.* **98**, 3247–3259 (1993).
52. Haberle, R. Interannual variability of global dust storms on Mars. *Science* **234**, 459–461 (1986).
53. Anderson, F. et al. Assessing the Martian surface distribution of aeolian sand using a Mars general circulation model. *J. Geophys. Res.* **104**, 18991–19002 (1999).

Acknowledgements

I thank M. Smith for making available Figs 2 and 5, and B. Cantor for making available Fig. 4.

Mars exploration

Michael H. Carr* & James Garvin†

*US Geological Survey, Menlo Park, California 94025, USA

†Code SR, NASA Headquarters, Washington DC 20546, USA

An international flotilla of spacecraft are to be sent to Mars over the next decade in an effort to understand the planet's geology and climate history, and to determine whether some form of life ever started there. At least two spacecraft will be sent at each launch opportunity, and at times up to four spacecraft may be operating simultaneously at the planet.

The exploration of Mars is an international endeavour involving many nations. The United States has recently reaffirmed its long-term commitment to Mars exploration, the European Space Agency is set to launch several spacecraft to the planet in the next few years, Japan has a spacecraft (Nozomi) *en route*, and other countries are making substantial commitments to the effort. This global interest stems from a variety of causes. Mars is the most Earth-like of the other planets in our Solar System. Like the Earth, it has had a varied geological and climatological history¹. Most of the planetary processes familiar to us here on Earth have apparently operated also on Mars, although under very different conditions, and perhaps at different scales and rates. Of particular importance is that Mars is the only planet in our Solar System, other than Earth, where liquid water has had a significant role in the evolution of the surface². This fact, together with indications that Mars' climate may have been more Earth-like in the past³, has led to speculation that some form of indigenous life could have developed there⁴, or that the planet may have been colonized successfully by terrestrial life as a result of interplanetary transfer of meteorites⁵. These possibilities have been further stimulated by the finding of enigmatic structures and mineralogical relations in martian meteorites, which have been interpreted as possibly of biological origin⁶. An additional stimulus to Mars exploration is that it will surely be the first planet outside the Earth–Moon system to be visited by humans.

Strategic issues

While the long-term prospects for Mars exploration may include manned missions, the present exploration strategy is robotic and largely science driven, although constrained, of course, by the practicalities of budgets, schedules and technical feasibility. Determination of many environmental factors relevant to human exploration are, however, encompassed in the science goals. In recent years the United States has adopted an exploration strategy informally referred to as faster, better and cheaper. It involves launching to Mars many small, relatively low-cost missions, with narrowly focused science objectives, rather than large, costly complex missions with comprehensive science goals. The intent has been to drive down costs and to distribute risk across several missions so that a single failure will not catastrophically disrupt the whole programme. The strategy was adopted by the US Mars Program after the loss of Mars Observer in 1993. Before this time, planetary missions had grown so costly and complex that only one mission could be launched each decade. In contrast, since adopting the faster-better-cheaper strategy, the United States has

launched five missions to Mars alone. The strategy has, however, had only mixed success. Although Pathfinder and Mars Global Surveyor were enormously successful, two Mars spacecraft were lost at Mars in 1999, probably because costs were so reduced that unacceptably high risks were taken. Despite the two failures, multiple, small missions still form the basis of the US strategic plan for Mars, although, in light of the failures, more attention is being paid to containing risks.

The orbital motions of Earth and Mars cause opportunities to launch spacecraft to Mars to recur every 26 months. The United States plans to launch at least one, and sometimes two, spacecraft to the planet at every opportunity for at least the next decade. In addition, the European Space Agency plans a launch in 2003 and France and Italy are planning to launch vehicles to the planet later within the decade, as described below. Thus, the exploration of Mars is truly international. Strategic planning meetings have multinational representation, most payloads are international, and some missions may be dependent on another agency for some essential element, such as telecommunications. Information on the various national plans is exchanged through the International Mars Working Group with representation from 11 nations.

The three basic science elements of Mars exploration are searching for evidence for past or present life, elucidating the evolution of the solid planet, and determining the climate history. Although the search for life is a primary objective, it entails high risk in that there may never have been any biological activity on the planet. Thus, the prudent strategy is, while pursuing the search for life, to continue to address the more general objectives concerning the evolution of the planet and its atmosphere. Mars is a very diverse planet with an area roughly the same as the land area of the Earth. Expectations are that if life ever evolved on the planet, or if there is life there today, the evidence will be preserved only in rare locations where conditions for fossil preservation or for sustaining life are met. The search-for-life strategy is, therefore, to first map the planet globally from orbit in an attempt to better understand what is where and to search for locations where water, warmth and nutrients may have been available to support life⁷. The global data will also enable us to reconstruct the broad outlines of the geological and climate history. The next step is to make more detailed observations of targets of interest both from orbit and on the ground. Finally, armed with the global and local knowledge, sites will be picked for comprehensive *in situ* life-detection experiments and sample-return missions. Meanwhile, experiments addressing objectives less directly concerned with life detection, such as achieving a better understanding of the circulation of the atmosphere or the structure of the

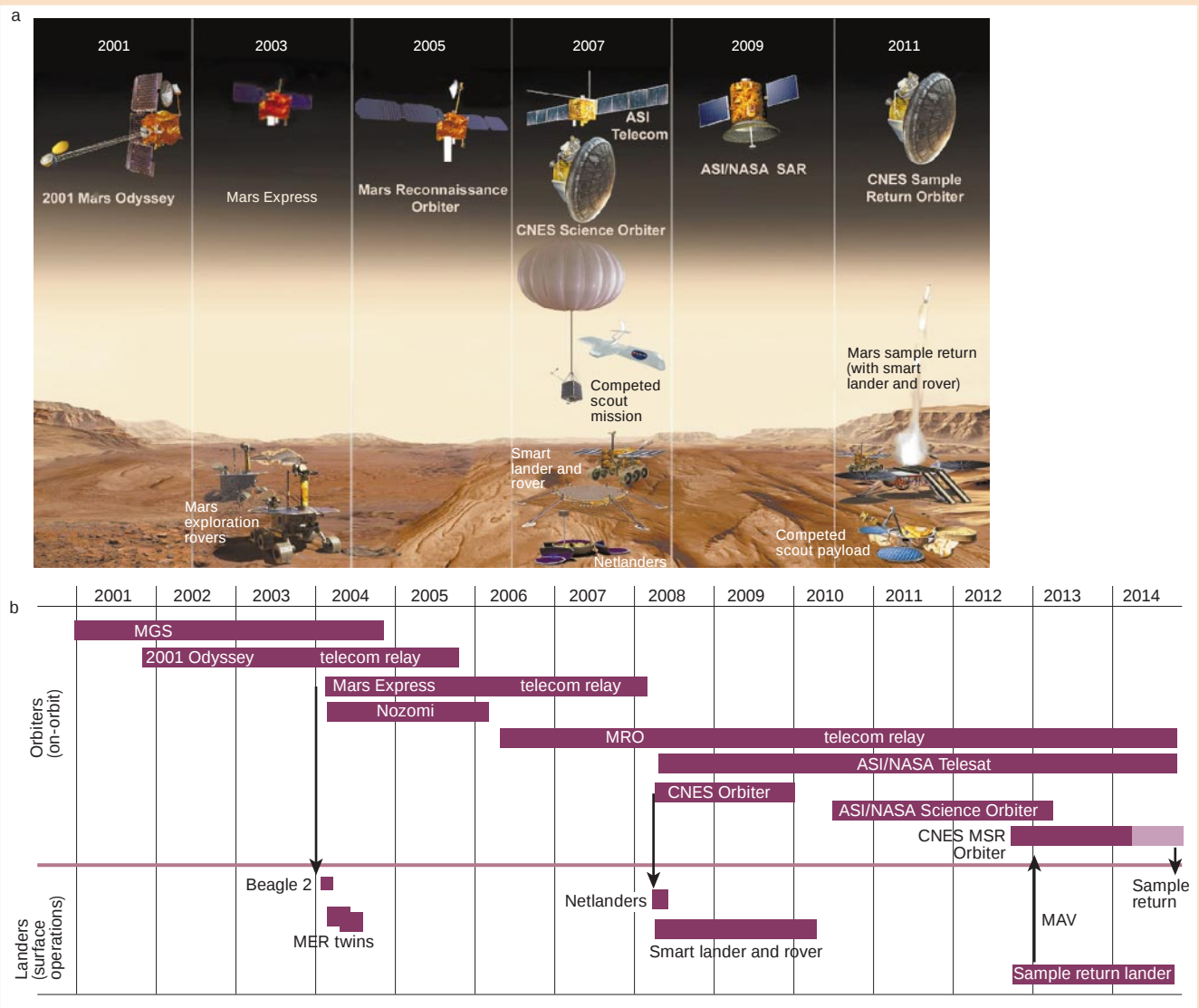


Figure 1 Envisaged mission sequence. **a**, Cartoon showing the different types of spacecraft to be launched to Mars in the next decade. The dates indicate the year the spacecraft is launched. **b**, The bars indicate the times that different spacecraft are expected to be operating at Mars (CNES and ASI are respectively the French and Italian space agencies).

interior, will be interleaved with the other missions. The strategy thus allows different high-priority investigations to be undertaken in parallel.

One of the main issues in the long-range strategic plan is when to return samples. Chemical and isotopic measurements can typically be done in terrestrial laboratories with precisions and sensitivities that are orders of magnitude better than are possible *in situ*. Other techniques, such as precise age determination and detection of microfossils, are so complicated and require such intricate sample preparation as to be impractical on a distant planet in the foreseeable future. Moreover, it is impossible to know *a priori* what are the most critical measurements to make *in situ*, and how best to make them. In contrast, having samples here on Earth allows the full analytical capability of the science community to be applied to them. The measurement strategy can shift in response to previous analytical results, and new techniques can be developed, as suggested by the samples themselves. Many researchers believe that the crucial questions of the age of surface materials, the nature of changes in climate, and whether life ever arose on the planet will not be answered definitively until we have here on Earth a well documented set of samples that are representative of the planet's variety. But sample return is expensive and technically challenging. Several new capabilities must be

developed, such as autonomous launch from the martian surface and rendezvous in Mars orbit. The costs of developing these capabilities must be spread over several years. Moreover, it is argued that our knowledge of Mars is currently insufficient to know where to go to collect the samples that are most likely to resolve the question of whether there has ever been life on the planet. Preservation of fossils or evidence of biological activity requires special conditions; if there is life today, it is likely to be found only in very local niches. Because we are likely to get samples from only a few places, it is argued, we should be choose them particularly carefully. As a result of these technological and scientific challenges, a sample-return mission is unlikely to be launched before 2010, despite its importance.

Another argument raised against an early sample return is that if life exists on Mars today it could present a hazard to terrestrial life on Earth⁸. Almost everyone who has examined this problem has concluded that the chances that returned samples present a hazard are extremely small. After all, numerous martian meteorites fall on the Earth every year. Nevertheless, we cannot prove that the probability that the samples are harmful is zero. Accordingly, any sample returned to Earth must either be sterilized or returned in such a manner that the risk of inadvertent release is extremely low. Sterilization undermines the rationale for the samples, as in the process the

materials of most interest are likely to be altered or destroyed. Thus, the focus on sample-return studies has been on designing systems to return samples to Earth in sealed containers that would be opened only in stringently controlled laboratory conditions⁹. Because of such issues of back contamination, some biologists have argued that, rather than returning samples, we should concentrate on developing sophisticated life-detection techniques that could be used at Mars, thereby by-passing the contentious problems of containment and hazard assessment.

The next decade

The next decade will be an active one for Mars exploration. As indicated above, missions will be sent to Mars at every 26-month opportunity, and on most opportunities there will be multiple launches. Figure 1 shows the currently envisaged mission sequence. The missions up to the 2005 launch opportunity are being implemented and, barring unforeseen circumstances, will be launched as indicated. Missions beyond 2005 are less certain. One consideration in the US strategy that was revised after the failures in 1999 was to have, where possible, orbiter missions succeed each other every four years, and likewise with lander missions. This was to enable lessons learned from one mission to be incorporated into a similar mission downstream. Thus, orbiter missions are to be launched by the United States in 2001, 2005 and 2009, whereas lander missions are to be launched in 2003, 2007 and 2011. The early non-US missions, having been started before the revised strategy, are interleaved with the alternating orbiter and lander missions.

Mars Global Surveyor was injected into Mars orbit in September 1997, and has been returning data ever since. It has imaged a few per cent of the surface at resolutions of 1.5–8 m pixel⁻¹ (ref. 10), determined elevations of the whole planet at a spatial resolution of a few hundred metres and a vertical resolution of better than 1 m (ref. 11), and obtained thermal infrared spectra of the entire surface with a spatial resolution of 3 km (ref. 12). It has also determined the gravity field to degree and order of roughly 100 (ref. 13) and discovered large crustal magnetic anomalies which indicate that early Mars once had a strong magnetic field¹⁴. The spacecraft is currently making detailed observations of potential landing sites for subsequent missions, and in particular the 2003 rovers. It also now seems likely that Mars Global Surveyor will survive long enough to act as a data relay for other missions.

The next Mars mission is Mars Odyssey (2001), which will be placed in Mars orbit in October 2001. It carries three experiments. A gamma ray spectrometer will map the elemental composition of the surface globally at a spatial resolution of 250–300 km. A thermal emission imaging system will acquire infrared spectral images with a spatial resolution of 100 m of locations on the surface of special interest. This experiment is of particular interest for choosing landing sites for future missions as it may locate chemical anomalies such as might be caused by hydrothermal activity. The third experiment, MARIE (for Mars Radiation Environment Experiment), is intended to characterize the galactic cosmic radiation environment around Mars, in anticipation of engineering design choices that must be made for future human missions to the planet. Odyssey is also designed to act as a relay for the rovers to be launched in 2003.

Several missions will be launched to Mars in 2003. The United States will launch two Mars Exploration Rovers. These will soft land on the surface in air bags like Mars Pathfinder did in 1997, but unlike Pathfinder, the rovers will be independent of the landing system and be free to move away from the landing point. The landers are expected to survive on the surface for at least 90 days, during which they will perform a variety of measurements on the local rocks and soils. The process of choosing and characterizing the landing sites is well underway. Candidates include the floor of the canyons, floors of large craters thought to have formerly contained lakes, the edge of the ancient heavily cratered highlands, and a region where Mars Global Surveyor detected coarse-grained haematite (an iron mineral

believed to have been precipitated from water) at the surface. Also in 2003, the European Space Agency is scheduled to launch an orbiter called Mars Express. This will carry a German high-resolution camera capable of mapping much of the planet in stereo and colour. It will also carry a French near-infrared spectrometer for determining surface mineralogy. In addition, a joint Italian/US radar surface sounder will attempt to detect liquid water down to depths of ~5 km below the surface. Other experiments will make measurements on the upper atmosphere and characterize the interaction between the planet and the solar wind. These plasma experiments will complement those of the Japanese spacecraft Nozomi, which will be in orbit around Mars at the same time. Mars Express will also deploy to the surface a small UK-built lander, called Beagle 2. It has a camera and a variety of detectors to measure the composition of the surface and atmosphere. Thus, 2004 will be a banner year for Mars exploration, with three landers on the surface and at least two and possibly four spacecraft in orbit, as the Odyssey and Mars Global Surveyor spacecraft could both survive that long and act as relay links.

In 2005, the US plans to launch a Mars Reconnaissance Orbiter. The payload for the orbiter has yet to be chosen, but it will probably include an atmospheric sounder designed to acquire multiple vertical profiles over one martian year, in order to characterize the atmosphere's circulation. Other likely instruments are a camera with resolution of tens of centimetres, and a visible-to-near-infrared spectrometer, both intended for use in identifying and characterizing future landing sites. There is also likely to be a radar sounder to detect subsurface water, but with a greater sensitivity and spatial resolution than that to be flown on Mars Express.

Beyond 2005, plans are much less certain. Likely possibilities include a 2007 launch of a French experiment to place four netlanders on the martian surface. Their main purpose will be to establish a seismic network to determine the internal structure of the planet, but they will also have imaging and geochemical sensors. At the same time, the United States plans to send additional rovers to the surface, to do longer-range exploration than those sent previously and to test some of the technologies needed for sample return. The United States will also request proposals from the science community for a small mission, termed a Scout, that could fill gaps left in the main programme. A joint US/Italian telecommunications satellite to support the various vehicles on the surface is also a possibility. What happens after 2007 will depend on budgets, on technological developments and also on what has been discovered about Mars in the intervening years. As indicated above, sample return is a high priority. By 2007 we should have acquired most of the orbital observations needed to select potentially fruitful landing sites, and have made enough measurements on the surface to provide the ground data required to interpret with confidence the orbiter remote-sensing data. We will also have considerable experience landing and operating vehicles on the surface. Thus, if we continue to maintain the 4-year spacing between landed missions, a 2011 launch of a sample-return mission is a possibility.

Beyond the next decade

What we do in the second decade of the millennium will depend on what we discover during the next few years, what funds are available, and whether we are any closer to committing to the exploration of the planet with people. Perhaps the most important factor governing what follows will be whether life is unambiguously detected either in martian meteorites, returned samples or by *in situ* experiments. Should life be detected, then the focus of subsequent exploration would surely be on characterizing that life and determining how and when it evolved. But the chance that life will be unambiguously detected is low. And whether or not it is detected, the desire for samples will remain high. We will want to acquire samples of different-aged igneous rocks, hydrothermal deposits, evaporites, lake beds, ancient channel sediments, deposits downstream of recently formed gullies, and so forth. The samples will be needed to further

the search for life and to acquire more definitive information on the timing and nature of geological and climatological events. Long-lived rovers that can travel hundreds of kilometres with sophisticated geochemical and life-detection payloads will no doubt be advocated in order to sample a wide variety of geological units and possible biological environments. Such missions could be coupled with a capability for returning samples so that we can obtain a suite of samples representative of the planet's variety. Suggestions have also been advanced to drill into the surface to reach depths where liquid water might be stable today, and where life might still survive, as it does deep within the Columbia River basalts in eastern Washington¹⁵. Drill cores from the polar layered terrains would be of particular interest for unravelling the recent climate history. Other possibilities that have been suggested include low-flying balloons and aircraft to detect local subsurface water and thermal anomalies, and deploying global seismic and meteorological networks. Most of these suggestions do not readily fit within the faster-better-cheaper mould. But if we are to have a long-term exploration programme, we will need larger, more complicated missions than are currently being flown.

A major issue in the long-term plan is the role of human exploration. Opinions differ markedly on the role of humans. Some argue that humans should not set foot on the planet until it has been demonstrated that there is no life on Mars today. Humans would inevitably be accompanied by terrestrial microbial life that might compete with any indigenous life, thereby destroying or altering it. Humans could also return martian microbes to Earth, and so put some terrestrial life at risk. Others argue that such concerns are overstated, that the chances of there being life on Mars today is close to zero, and that even if there is life and it is brought back to Earth, it could not compete with our planet's terrestrial organisms, which occupy every conceivable niche. Another issue is what purpose it would serve to send people to Mars. The enormous cost of transporting people simply to perform science experiments and fieldwork

hardly seems justifiable. For the cost of one manned mission, hundreds of robots could be deployed all over the planet to do fieldwork, collect samples and deploy instruments. Some advocates of manned missions assert that the rationale for human missions to Mars is really not practical but spiritual. The will to explore is deeply rooted in human nature. Exploration lifts us above the humdrum concerns of food and shelter and provides us with feelings of awe, wonderment and pride. If people are ultimately to go to Mars, it may well be such considerations that drive us there. But human exploration is a long way off. Meanwhile, there is an entire planet to explore and possibly an entirely new biology to discover. □

1. Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S. (eds) *Mars* (Univ. Arizona Press, Tucson, 1992).
2. Carr, M. H. *Water on Mars* (Oxford Univ. Press, New York, 1996).
3. Haberle, R. M. Early Mars climate models. *J. Geophys. Res.* **103**, 28467–28479 (1998).
4. McKay, C. P., Mancinelli, R. L. & Stoker, C. R. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 1234–1245 (Univ. Arizona Press, Tucson, 1992).
5. Gladman, B. J., Burns, J. A., Duncan, M., Lee, P. & Levison, H. F. The exchange of impact ejecta between terrestrial planets. *Science* **271**, 1387–1392 (1996).
6. McKay, D. S. *et al.* Search for past life on Mars: possible relict biogenic activity in martian meteorite ALH84001. *Science* **273**, 924–930 (1996).
7. National Aeronautics and Space Administration. An Exobiological Strategy for Mars Exploration. Report No. NASA SP-530 (NASA, 1995).
8. National Research Council. *Mars Sample Return. Issues and Recommendations* (National Academy Press, Washington DC, 1997).
9. National Aeronautics and Space Administration. Mars Sample Handling and Requirements Panel Final Report. Report No. NASA/TM-1999-209145 (NASA, 1999).
10. Malin, M. C. & Edgett, K. S. Mars Orbiter Camera: the first three years. *J. Geophys. Res.* (in the press).
11. Smith, D. E. *et al.* The global topography of Mars and implications for surface evolution. *Science* **231**, 1495–1503 (1999).
12. Christensen, P. R., Bandfield, J. L., Smith, M. D., Malin, V. E. & Clark, R. Identification of a basaltic component on the martian surface from Thermal Emission Spectrometer data. *J. Geophys. Res.* **105**, 9609–9621 (2000).
13. Zuber, M. T. *et al.* Internal structure and early thermal evolution of Mars from Mars Global Surveyor topography and gravity. *Science* **287**, 1788–1792 (2000).
14. Acuna, M. H. *et al.* Magnetic field and plasma observations at Mars: initial results of the Mars Global Surveyor mission. *Science* **279**, 1676–1689 (1999).
15. Stevens, T. O. & McKinley, J. P. Lithoautotrophic microbial ecosystems in deep basalt aquifers. *Science* **270**, 450–454 (1995).

MARS EXPLORATION

At The Planetary Society, We Make it Happen!

For two decades, The Planetary Society has helped to advance the scientific exploration of Mars. Through research support, political advocacy, and educational action, we seek to fulfill our mission to inspire the people of Earth to explore new worlds and seek other life.



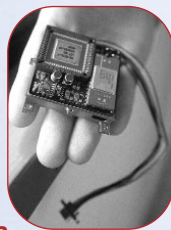
Mars Balloon—

It's an elegant idea whose time has not yet come, but one that holds promise for future exploration. Working with the French and Russian space agencies, The Planetary Society developed and tested an instrumented guide-rope, carried by a balloon, that would enable low-altitude flight on Mars and return scientific data from the surface.



Mars Rover—

Before international cooperation in space science received governmental blessing, The Planetary Society brought together scientists and engineers from around the world to test robotic technologies for exploring Mars. From the volcanic slopes of Kamchatka to the rocky bottom of Death Valley, we've driven rovers across Earth's harshest terrains, gaining experience we hope to apply someday to Mars.



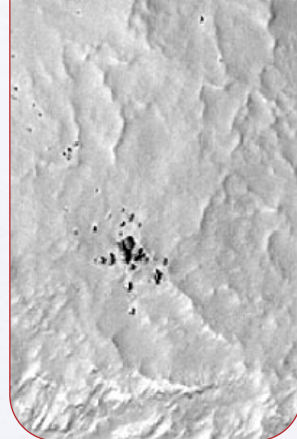
Mars Microphone—

It was simple and inexpensive, would capture the public's attention, and we couldn't resist the opportunity to do it. The Planetary Society created and funded the microphone that flew on the Mars Polar Lander. Although that craft was lost, we're now preparing a suite of microphones to fly on CNES's Mars Netlander mission.



Highway to Mars?—

We are working toward an expansive future when humanity will boldly explore other worlds. The Planetary Society is set to launch Cosmos 1, the first solar sail, late this year. Solar-sail-powered craft could someday cruise a highway between Earth and Mars, carrying supplies for future explorers of the Red Planet.



Red Rover Goes to Mars—

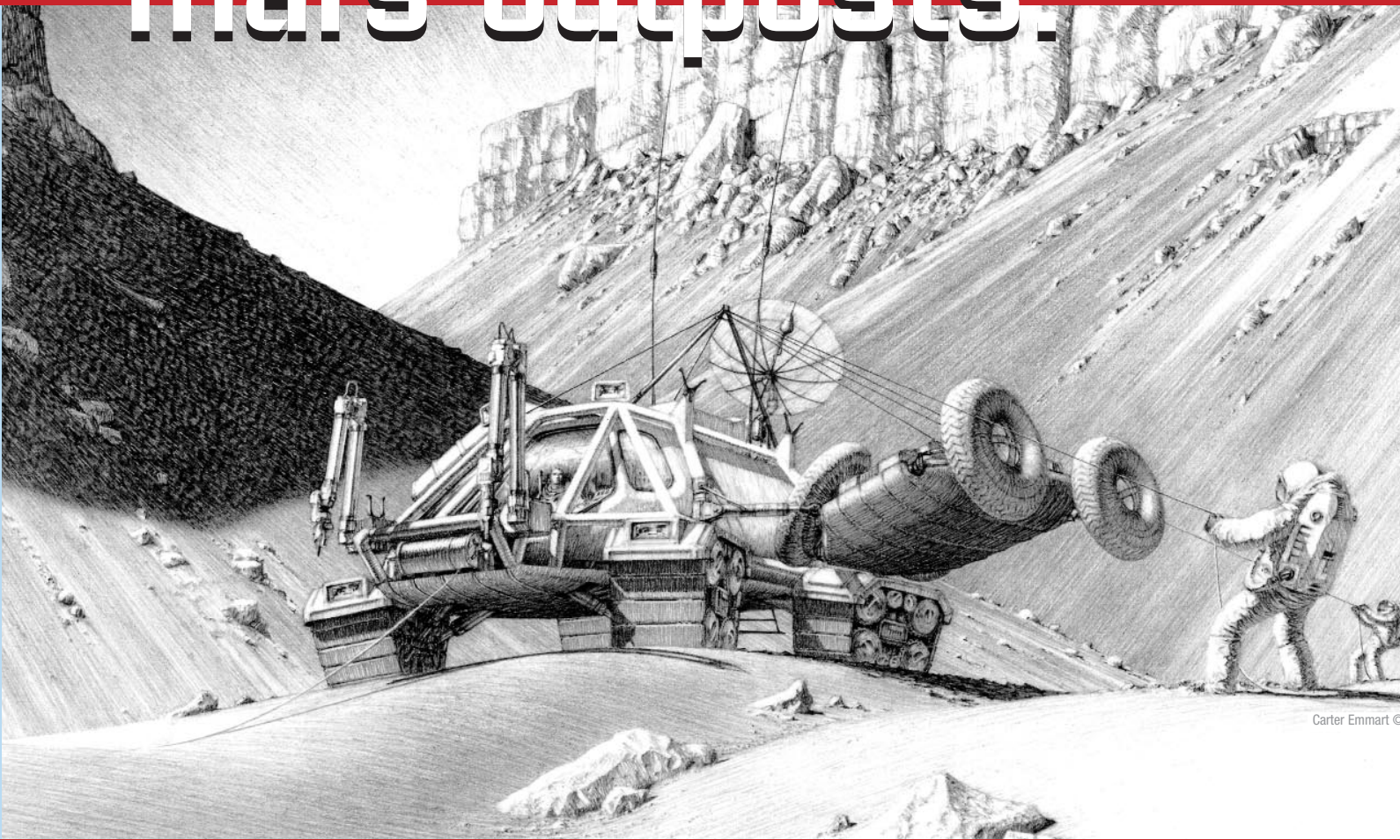
Scientists and engineers of the future are getting a chance to become true participants in real missions to Mars. The winners of the first leg of our Red Rover Goes to Mars contest recently used the camera on the Mars Global Surveyor to discover a group of black boulders whose origin and placement has puzzled experienced planetary scientists. Future winners will actually help control the Mars Exploration Rovers set to land in 2004.



MARS BECKONS. HELP US EXPLORE. JOIN THE PLANETARY SOCIETY TODAY.

PLANETARY.ORG

Mars Outposts:



A Planetary Society Approach to Exploration

by Bruce Murray, Wesley T. Huntress, Jr. and Louis D. Friedman

We are transported to the planet's surface. The temperature is -40 degrees Fahrenheit—a typical day. Wispy clouds skirt across the thin, reddish-ocher sky. We navigate over a rocky slope and peer for the first time into the mouth of a ravine, its walls apparently shaped by flowing water. The erosion seems relatively recent. Could water still be percolating to the surface, then evaporating? Is there evidence of life below the surface? We press on for clues.

All this can become reality within the next decade.

Bridging the Gap

Until now, exploration has involved either robotic spacecraft or human expeditions. We launch probes to distant worlds to collect scientific data. Closer to home, humans regularly travel to Earth orbit and have ventured to the Moon.

Robotic spacecraft are relatively inexpensive, costing, on average, several hundred million dollars. By contrast, human exploration is extremely expensive. To place the first humans on the Moon consumed, at one point, about 5 percent of the annual US federal budget and took nearly a decade to complete. The price tag for a human expedition to Mars—a goal that generates broad public interest—is many tens of billions of dollars and could take at least another decade.

Whether by robotic or human exploration, Mars beckons.

Rather than argue about which approach is preferable, a third way bridges the gap between robotic and human exploration. The Planetary Society has dubbed this approach Mars Outposts.

Mars Outposts would consist of specially designated research sites on the Red Planet, equipped with permanent communications, navigational systems, and other technologies to support intensive robotic missions and, most important, vicarious public participation. The areas would become, in a sense, “Martian Antarctica,” places of high scientific interest where researchers from around the world collaborate to learn about the planet.

At the sites, rovers, balloons, and other probes would comprehensively investigate the surrounding terrain. Thanks to continuous signals broadcast to Earth and distributed through the Internet, humans worldwide would be able to participate in the exploration of the planet. For instance, via camera lenses on a rover, students could explore the landscape and command the vehicle, maneuvering through dry river basins and along the polar regions. Mars is a magnificent place. A deep swath cutting across the face of Mars dwarfs America's Grand Canyon. Olympus Mons rises more than twice the height of Mount Everest.

With Mars Outposts, the whole world could collectively experience the thrill of exploring another world. We would all become the Lewises and Clarks of Mars.

Step by Step

Furthermore, the Mars Outposts approach incrementally establishes the infrastructure needed for human expedition and thus greatly reduces costs and increases safety. Also, the outposts could become future landing sites. The same communications and navigational systems used for the robotic probes could later support a human mission. The robotic infrastructure, for instance, could facilitate the production and storage of propellant and also breathable oxygen produced on Mars from the planet's thin, carbon dioxide atmosphere—thereby reducing the payloads launched from Earth to support initial robotic sample return and subsequent human exploration and return to Earth.

Additionally, the outposts would allow scientists and engineers to determine the tasks best accomplished by robotic technologies and those more appropriately performed by humans. The new paradigm of human-machine symbiosis could be built step by step.

Humans will travel to Mars; we just don't know when. The next step to realizing this dream could be aboard the International Space Station—there conducting research on keeping humans healthy in space over extended periods and preparing them for the long and arduous expedition to Mars. The station also could be used to develop habitat technology for a Martian crew.

As the space station is an international endeavor, so, too, will be a human expedition to Mars. And so, too, would be the Mars Outposts. Together, all would share the costs, risks, and opportunities of robotically learning about Mars and progressing to the human phase. The outposts would catalyze the transition to direct probing of the habitability of Mars.

It is a propitious time for spacefaring nations to initiate the Mars Outposts program. New discoveries from Mars Global Surveyor, now orbiting the Red Planet, strongly suggest evidence of the presence of water and thus possible outposts.

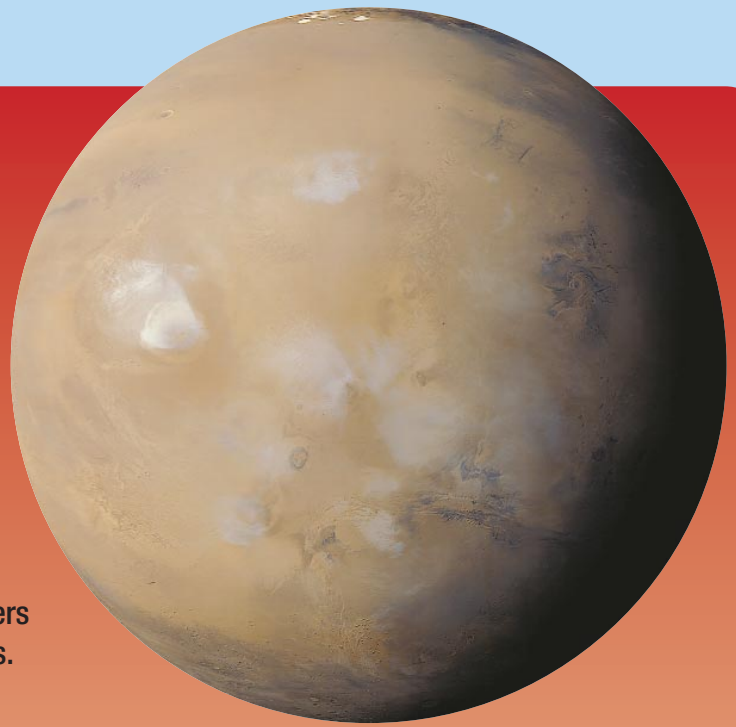
To begin the process, outposts would be selected on Mars; then promising sites would be expanded at an affordable level. Over time, these areas would become focuses of scientific research and inspirations for a generation. We are blessed to live at a time when we are able not only to dream about distant worlds but also to explore them together. Ours is the first generation of the Space Age. It can be our fortune to reach into the heavens as one and open our eyes upon a new world.

Bruce Murray is President, Wesley T. Huntress, Jr. is Vice President, and Louis D. Friedman is Executive Director of The Planetary Society.

THE PLANETARY SOCIETY: WE MAKE IT HAPPEN!

Founded in 1980 by Carl Sagan, Bruce Murray, and Louis D. Friedman, The Planetary Society has, for more than two decades, led the charge to uncover the tantalizing secrets of the Red Planet.

Representing more than 100,000 members from more than 140 countries, The Planetary Society is the one organization uniquely poised to advocate strong space programs worldwide. Our voice is heard by NASA, by the European Space Agency, and by leaders of space programs in Japan, Russia, and other nations.



**BECOME A MEMBER OF THE PLANETARY SOCIETY AND ADVANCE TO THE FOREFRONT
OF THE EXPLORATION OF MARS AND BEYOND.**



Be Part of the Adventure—Join The Planetary Society Today!

<http://planetary.org>

1-800-9WORLD5 (toll-free within the US)

1-626-793-5100 (international)

The Planetary Society • 65 North Catalina Avenue • Pasadena, CA 91106-2301 • USA

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:

Leonie Welss (4954)

European Manager:

Nevin Bayoumi (4978)

European Sales Consultant:

Jonathan Bull (4973)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Silje Opstrup (4994)

Production Manager:

Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Amanda Wren
Tel + 34 91 391 3835
Fax + 34 91 319 0091
e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: classified@natureny.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi Shinjuku-Ku,
Tokyo 162
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejp.com
Japan Manager: Kate Cowan

Prosperous physicists

Things are looking up for US physicists — at least for now. A recent survey (<http://www.aip.org/statistics>) carried out by the American Institute of Physics (AIP) among its members shows that unemployment has gone down and that salaries have significantly increased since the last survey in 1998. Unemployment among respondents was at 0.6% in 2000, the lowest in a decade. Respondents working in industry, medical services, university-affiliated research institutes and universities with 9–10-month contracts reported median salaries 5–14% higher than the median levels in 1998.

But salaries varied widely by sector and sub-discipline. For example, AIP members working in hospitals or medical services earned a median income of \$100,000, whereas respondents working in 4-year colleges on 9–10-month contracts earned half that figure.

As one might expect, the good news does not come without some qualifications. Because the survey only covers AIP members, the results could be somewhat skewed. Presumably, the survey did not include any disgruntled physicists-turned-cab-drivers who let their membership lapse.

Overall, any gains at all in any sub-population of US physics is cause for celebration. But US physicists may want to be cautious about celebrating too much, because the past decade has shown that gains can be fleeting.

The present climate is no exception. President George W. Bush's proposed budget favours tax cuts over new investments in domestic programmes. Still, physicists funded by the Department of Defense and the National Institutes of Health should face continued good fortune because those agencies receive strong support from both the president and Congress.

Paul Smaglik

Naturejobs editor



Contents

MOVERS

Theoretical physicist heads for Singapore; protein researcher takes on R&D challenge; and more

[Back page](#)

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



ASTROPHYSICS

Günther Hasinger, who helped to build up the Astrophysical Institute Potsdam (AIP) after German unification, will direct high-energy astrophysics at the Max Planck Institute for Extraterrestrial Physics in Garching starting in October. Hasinger now serves as managing scientific director of the AIP, which he joined in 1994. He enjoyed the management challenge of bringing the AIP, which was fairly isolated before unification, into contact with more researchers in the West, including the United States. But he looks forward to returning to his research interests in high-energy astrophysics, particularly with X-ray telescopes. At Garching, Hasinger will lead groups in building and using instrumentation for space-based observations. "I could not have done this at the same level in Potsdam," he says, noting that Garching focuses more on technology development and instrumentation.

BIOINFORMATICS

Gunaretnam Rajagopal, a theoretical physicist formerly affiliated with the University of Cambridge, took the helm of the Singapore Bioinformatics Institute this month. Gunaretnam, who has a strong foundation in supercomputing and a keen interest in biology, will focus the institute on training about 100 postgraduate students a year to help to meet the demand for computing experts who can make sense of the information from the genomics revolution. As well as providing training, the institute will support projects including the fugu fish genome-sequencing project, now being carried out by the Institute of Molecular and Cell Biology, the US Department of Energy and the Singapore Genomics Programme. Gunaretnam, who is Malaysian, says he will be happy to be closer to his mother, who lives in Kuala Lumpur.

CHEMISTRY

LGC, the United Kingdom's leading independent analytical laboratory (formerly known as the Laboratory of the Government Chemist), has appointed **Ian Kent** as chairman. The former chairman of Roslin Bio-Med will take over at LGC when Brian Richards retires from the board. Kent has spent his entire career managing agrochemical companies and is non-executive chair of several firms besides LGC. He also has a background in biotechnology as a founder of Imutran, one of the first xenotransplantation companies, which was sold to Novartis in 1996. Kent aims to help to continue the expansion of the once-public LGC as

a private company. The company provides chemical- and DNA-based analysis, including verification of analytical methods used by chemical and pharmaceutical companies. It also processes samples for the United Kingdom's criminal forensic DNA database. LGC is expanding its activities, including the growth of its reference materials business with the June acquisition of Promoshem. "The big issue is managing growth, which is a nice problem to have," says Kent.

PHARMACEUTICALS

Peter Kim's path from the Massachusetts Institute of Technology's Whitehead Institute to executive vice-president, research and development, with Merck marks a logical pursuit of a scientific quest as well as a natural career progression. Kim joined the Whitehead as a postdoc in 1985, became an associate member in 1988, and was named a Howard Hughes Medical Institute investigator in 1992. At the Whitehead he concentrated on protein folding, and in 1993 discovered, with Chave Carr, how a spring-loaded protein mechanism helps the influenza virus to invade targeted cells. They also learned that HIV, among other viruses, uses a similar mechanism. Kim soon began experimenting with peptides that might be able to inhibit viruses and he helped to form the biotech company Scriptgen in 1993. "I became very aware of how hard it is to make a drug," he says. That realization informed his decision to join Merck full time this February. "If my research could contribute to a drug or vaccine, that would be something I would consider one of the major accomplishments of my career," Kim says.

BIOTECHNOLOGY

Last month, HealthSpan Sciences named **Bryant Villeponteau** as its chief scientific officer. Villeponteau started his career in academia, first as an assistant research chemist in the Department of Chemistry and Biochemistry at the University of California, Los Angeles, then as assistant professor of biological chemistry at the University of Michigan Institute of Gerontology. During this period, Villeponteau was one of the first scientists to clone a human ageing-related gene. In 1993, Villeponteau joined Geron Corporation based in Menlo Park, California, then a start-up biotech company focused on age-related diseases, where he developed a patented technique for identifying gene-expression changes brought about by ageing. In 1994, Villeponteau and colleagues cloned and characterized the RNA component of human telomerase.



Günther Hasinger



Ian Kent



Peter Kim

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.